THÈSE DE DOCTORAT DE L'UNIVERSITÉ MARIE ET LOUIS PASTEUR

Reverse-correlation modeling of deficits of prosody perception in right-hemisphere stroke

Par *Aynaz ADL ZARRABI*

Thèse de doctorat de l'ED 37: Sciences Physiques pour l'Ingénieur et Microtechniques (SPIM), dirigée par *Jean-Julien Aucouturier* et *Marie Villain*

Soutenue à Besançon le *09 Juillet 2025*

Devant un jury composé de:

M. Jean-Julien AUCOUTURIER	DR CNRS, FEMTO-ST, Besançon	Directeur de thèse
Mme Anahita BASIRAT	PU, SCALAB, Univ. Lille	Rapportrice
Mme Charlotte JACQUEMOT	DR CNRS, IRMB/ENS, Paris	Examinatrice
M. Ladislas NALBORCZYK	CR CNRS, LPL, Aix en Provence	Examinateur
M. Emmanuel PONSOT	CR CNRS, STMS, Paris	Invité
M. Léo VARNET	CR HDR, CNRS, LSP/ENS, Paris	Rapporteur
Mme Marie VILLAIN	MCU HDR, Sorbonne Univ., Paris	Co-encadrante

Acknowledgment

First and foremost, I would like to thank my PhD supervisors, Jean-Julien and Marie. They made something that seemed so difficult to pursue possible for me, especially coming from a very different background and not knowing much about neuroscience, linguistics, or clinical cases at the start.

Marie, thank you, if today I know anything about clinical cases and prosody, it's mostly grâce à toi. You connected me with collaborators, including the Master's student and the ICM team, and I truly appreciate it. Jean-Julien, thank you for being such a passionate advocate for open science and for encouraging me to build toolboxes. Without your vision, this thesis would probably be much shorter! Your energy and clarity pushed me to improve constantly. What I learned from you, I will carry with me for the rest of my life. I am deeply grateful to have had both of you as mentors; such caring, supportive people who were always present, even with their busy workloads.

I would like to thank Dr. Anahita Basirat, Dr. Léo Varnet, Dr. Charlotte Jacquemot, and Dr. Ladislas Nalborczyk for agreeing to assess my PhD thesis. I look forward to the scientific discussions that will come from it. I also want to extend special thanks to Dr. Emmanuel Ponsot and Prof. Lionel Naccache for their enriching discussions on the ongoing project before my arrival, which resulted in a published work in Chapter 4. I am very grateful for the wonderful collaborative project with Dr. Ladislas Nalborczyk, which led to Chapter 6. I also thank Dr. Nathan Faivre for assessing the ongoing work during the thesis follow-up committee.

I am grateful to all my colleagues at the Neuro Group and at FEMTO-ST for their support; Paige, Rudradeep, Manoj, Zhenxing, Camille, Anaïs, and Coralie. I am also thankful that throughout this thesis, I had the chance to work and discuss with so many insightful people in particular Dr. Zeina Al Masry, Dr. Noura Dridi, Dr. Etienne Thoret, Dr. Karim N'Diaye, Zoé Rany, Dr. Annabelle Merchie, Célia Chauche and Aude Warney.

This journey would not have been possible without the support of my beloved family and friends. To my father, who inspired me to pursue my studies all the way to a PhD

and helped me become the person I am today. To my mother, who gave me the courage to be a fearless Iranian woman and stood by my side in every second of my life. This work is dedicated to my parents.

To my brother and sister, Joubin and Elnaz, my immigration journey would not have been possible without your help and love. I am also incredibly grateful to my friends Ghazaal, Ghazal, Shakiba, Asal, Magalie, Léa, Lisa, Adam, Seb, and Laurie; some of you from very far away, but always close to my heart. You stood by my side, encouraged me whenever I was down, and gave me the motivation to continue. You are all truly pearls.

Lastly, I would like to thank Alex, my life partner, who has played the biggest part in this achievement. I don't know how to thank you enough for your support, your belief in me, and your unwavering positivity that always gave me the courage and strength to go on. I look forward to all the journeys we will share together in life.

I am truly happy to have had the opportunity to fulfill my lifelong wish of continuing my graduate studies in France, a wish I worked so hard for, in my always beloved country, Iran, despite all the political issues, difficult circumstances, and the many limitations faced by women. As I close this chapter, my hope is that my compatriots may also achieve their goals and live the lives they aspire to. May their perseverance and dreams be met with the opportunities and freedom they deserve.

Contents

CHAPTER 1

Introduction

Beyond words themselves, our speech carries para-verbal information that conveys attitudes, emotions and intentions, often through prosody – the so-called melody of speech (Côté, Payer, Giroux, & Joanette, 2007). Prosody serves a variety of communicative functions, including emotional prosody (e.g., expressing happiness, anger or sarcasm through tone) and linguistic prosody (e.g., marking emphasis, syntactic boundaries or distinguishing questions from statements). Elements such as pitch/intonation, rate/speech rhythm, and timbre/articulation play a crucial role in expressing both emotional and linguistic meanings. However, following a brain stroke, particularly one affecting the right hemisphere, the ability to perceive or produce prosody can be significantly impaired, hindering effective communication (Etchepare & Prouteau, 2018). Although up to 54% of patients with right hemisphere damage exhibit deficits in prosodic comprehension (Blake, Duffy, Myers, & Tompkins, 2002), these impairments are often subtle and less noticeable than aphasia or motor dysfunction, resulting in underdiagnosis and insufficient treatment. Existing diagnostic tools, such as the Montreal Battery for the Evaluation of Communication (MEC; Joanette, Ska, and Côté (2004)), provide simple thresholds of performance but lack the sensitivity, specificity, and depth required to uncover the cognitive mechanisms

underlying prosody deficits (Benedetti, Weill-Chounlamountry, Pradat-Diehl, & Villain, 2022).

The goal of this thesis is to improve the diagnosis and comprehension of deficits of prosody perception after a brain stroke by capitalizing on an emerging psychophysical technique : reverse correlation (Neri, Parker, & Blakemore, 1999).

When studying the neural mechanisms that relate physical stimuli to perception, the modern field of psychophysics has indeed largely moved from simply measuring sensory thresholds and psychometric functions and now provides a toolbox of techniques to measure and fit multi-staged models able to simulate participant behaviour. Notably for the example of speech prosody, the psychophysical technique of reverse-correlation (or "classification images"; Murray (2011)) allows estimating, at the individual level, what sensory representations subtend the normal or abnormal perception of, e.g., interrogative prosody (Ponsot, Burred, Belin, & Aucouturier, 2018).

This thesis consists of 10 chapters, organized into 4 parts. **Part I** provides theoretical foundations for this work. **Chapter 2** introduces the biological foundations for what speech prosody is, for prosody perception deficits after right-hemisphere (RH) stroke and outlines the limitations of current clinical assessment tools. In **Chapter 3** we describe the history of psychophysical experiments and reverse-correlation method and explain how it can be used to extract Internal representation (from the cross-correlation of response and stimuli) and estimate internal noise, which reflects the response consistency in these experiments.

Part II provides a first reverse-correlation encounter with clinical data, leading to the definition of the thesis' specific problem statement. **Chapter 4** presents the application of reverse correlation to model differences in linguistic prosody processing between healthy participants and RH stroke survivors and introduces new behavioral biomarkers of prosodic processing that also correlate with clinical measurements. This work was conducted in collaboration with Mélissa Jeulin, Emmanuel Ponsot, Pauline Bardet, Pauline Commère, Lionel Naccache, JJ Aucouturier and Marie Villain and was published in Scientific Reports journal in 2024 as "A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after

right-hemisphere stroke" (Adl Zarrabi et al., 2024). Based on this initial analysis, **Chapter 5** identifies a number of limitations of existing state-of-the-art techniques for analyzing reverse-correlation data, which make them unsuitable for our patient population, who differ from controls by their fatigability, low consistency, and tendency to perseverate in their responses.

Part III introduces the main methodological contributions of the work. **Chapter 6** introduces three novel internal noise estimation methods that address the limitations identified in Chapter 5. As the time of submitting this manuscript, we are preparing a paper on this work in collaboration with Ladislas Nalborczyk , JJ Aucouturier and Marie Villain. In **Chapter 7**, we then introduce a novel kernel estimation procedure based on the GLM-HMM architecture (Ashwood et al., 2020) which is able to estimate states of perseveration.

Finally, in **Part IV**, we apply these new techniques for a re-analysis of our clinical dataset. In **Chapter 8**, we study how new estimates of reverse-correlation parameters agree and complement existing gold standards for the diagnosis of prosodic impairment. Finally, in **Chapter 9**, we provide an exploratory analysis of what factors may influence perseveration in patients.

Last chapter concludes the thesis by summarizing the key findings, reflecting on their broader implications, and suggesting future directions for both clinical applications and methodological advancements.

This work has been supported by a grant from Fondation pour l'Audition (FPA RD 2021-12).

Part I

Theoretical foundations

CHAPTER 2

Biological foundations: Prosodic deficits after a right-hemisphere brain stroke

This thesis focuses on providing a new clinical assessment tool for a specific communication disorder affecting brain stroke survivors: *aprosodia*, or the inability to process the melody of speech. In part I, we provide some of the theoretical and methodological foundations for this work. In this first chapter, we provide some of the biological context to appreciate the construct of speech prosody (section 2.1), the clinical context of function disorders after a brain stroke (section 2.2), and the specific assessment of aprosodia (section 2.3). The next chapter will focus on providing methodological foundations for our proposed approach to improving the assessement of aprosodia, namely psychophysical reverse correlation.

2.1 Speech Prosody

2.1.1 Components of prosody

In spoken communication, how something is said can be just as important as what is said. Often referred to as the "melody of language" (Hellerman, 2003), prosody encompasses the song-like vocal modulations that shape meaning, express emotion, and clarify intent beyond the literal content of speech. For example, pauses can disambiguate syntactic structure (e.g., "the dog went into the kennel" vs. "the dog, lying in the kennel"), while changes in intonation can signal interrogative versus declarative intent (e.g., "he's coming tomorrow" vs. "he's coming tomorrow?").

In linguistic terms, prosody typically refers to the suprasegmental (i.e., spanning several phonemes) features of speech: rhythm (Cummins, 2000), intonation (Bolinger, 1958), intensity, and duration. It overlays phonemic content with information that modulates meaning and convey emotional or structural nuance. Additional features such as jitter (pitch variation) and shimmer (amplitude variation) contribute to voice quality, and may or may not be considered prosodic based on whether they correspond to pathological speech (Kreiman, Gerratt, & Gabelman, 2002) or have expressive function (Anikin, 2020).

Intonation (Figure 2.1) is perhaps the most salient acoustic cue in prosody and is defined by both its absolute height (pitch, which is the psychological correlate of fundamental frequency/f0) and its contour (e.g., rising, falling, or level pitch). These variations are detected early during auditory processing (Tang, Hamilton, & Chang, 2017) and underlie the perception of melody in both speech and music (dong Wang, Wang, & Chen, 2013; Plack, Oxenham, Fay, & Popper, 2005). In tonal languages like Mandarin, pitch contours directly alter word meaning; in non-tonal languages like French or English, they shape intonation patterns crucial for interpretation (Ponsot, Burred, et al., 2018). In this thesis, we will operationalize the processing of speech prosody primarily by looking at the perception of pitch contours.

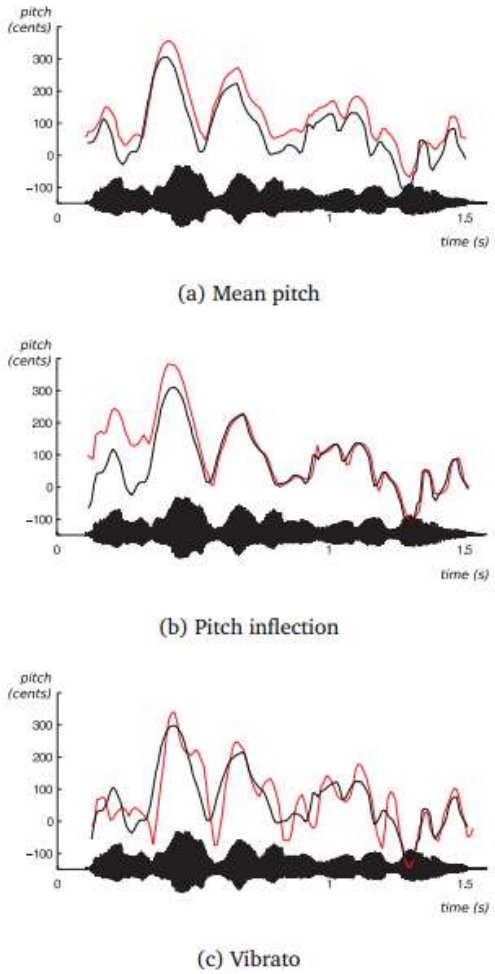


Fig. 2.1 **Illustration of variations of intonation that can be applied to convey a single sentence (French phrase: *"je suis en route pour la réunion"* / *"I'm on my way to the meeting"*).** **a** constant upward shift in mean pitch, for instance associated with positive emotion; **b** pitch inflection with an initial rise fading after 500 ms, for instance associated with high emotional arousal; **c** pitch oscillations (vibrato), typically signaling negative arousal such as anxiety or fear. Figure adapted from Rachman et al. (2017) where the variations were algorithmically generated. Red lines indicate the prosodically modified pitch contours.

2.1.2 Functions of prosody

The production of prosodic cues has several distinct functions, which are often studied in distinct research communities.

2.1.2.1 Linguistic Prosody

By providing information on top of the purely lexical sentence level, prosody can first play a crucial role in signaling grammatical structures, emphasizing contrastive elements, and delineating phrase and sentence boundaries. This function of prosody, often called *linguistic*, helps listeners interpret syntax, identify stress patterns, and differentiate sentence types (Baum & Dwivedi, 2003; Ross, Thompson, & Yenkosky, 1997). For example, the sentence "She didn't steal the money" can shift in meaning depending on which word is stressed: "**She** didn't steal the money" (implying someone else did) versus "She didn't steal the **money**" (suggesting she took something else). In tonal languages such as Mandarin Chinese, pitch variation determines lexical meaning: the syllable ma can mean "mother", "hemp", "horse", or "to scold", depending on the tone used (Howie, 1976). In contrast, in non-tonal languages like English or French, intonation patterns provide crucial grammatical cues—transforming "You're coming" into a question by simply raising pitch at the end (Ukaegbe et al., 2022). In this thesis, we will operationalize the processing of speech prosody by looking specifically at one of its linguistic function, namely its ability to convey questions using the final pitch rise (Banuazizi and Creswell (1999) and Fig. 3.6 in Chapter 3).

2.1.2.2 Affective Prosody

By providing several ways to pronounce the same words, prosody can also convey affective states, speaker intentions, and social attitudes (Ekman, Sorenson, & Friesen, 1969; Kamiloğlu, Fischer, & Sauter, 2019). This function of prosody, often called *emotional* or *affective* enables for instance the expression and recognition of joy, sadness, anger, irony, or sarcasm, critical elements of everyday social communication. For example, the phrase "I can't believe you did that" may express astonishment, anger, or admiration depending on the speaker's tone. Happiness is typically marked by increased pitch, faster tempo, and greater intensity, whereas sadness involves slower speech, reduced pitch, and lower amplitude (Belyk & Brown, 2014a). Unlike linguistic prosody, which is language-specific and governed by syntactic rules, emotional prosody is more biologically grounded and

culturally universal, with shared acoustic markers across languages and societies (Sauter, Eisner, Ekman, & Scott, 2010). Note that much of what is typically studied under the name of "affective" prosody does not encompass emotions stricto sensu: for instance, propositional attitude such as being critical, impressed or disapproving, or speaker behaviour such as being condescending, friendly or rude. Wichmann (2000) provides an excellent classification of such functions and how they relate to affective prosody.

Importantly, while linguistic and emotional prosody are often studied as distinct acoustic, cognitive and neurological phenomena, they often manifest themselves in identical physical features. For instance, rising pitch at the end of a sentence may have linguistic function (e.g., marking it as a question), emotional function (e.g., signaling the speaker's surprise), or express other speaker attitudes (e.g., uncertainty about the truth value of the sentence Goupil, Ponsot, Richardson, Reyes, and Aucouturier (2021a)), and perhaps all three simultaneously. Depending on context or attentional focus, they may or may not trigger certain cognitive processing, and there is much debate how discrete and independent these two functions really are (Belyk & Brown, 2014b; Seddoh, 2002).

2.1.3 Cerebral lateralization of prosody

While early theories proposed a right hemisphere specialization for affective prosody and a left dominance for propositional language (Ross, 1981), more recent work suggests a more complex and overlapping functional organization. Some researchers, such as Blumstein and Cooper (1974), argued that the RH processes all suprasegmental features, while the LH manages segmental language elements. Others, like Lancker (1980) and Behrens (1989), emphasized functional distinctions between the hemispheres, such as the RH's role in processing global sentence-level prosody, and the LH's role in local syllable-level structures.

Adding a temporal dimension to these anatomical and functional accounts, Schirmer and Kotz (2006) proposed a three-stage model of emotional voice processing (Figure 2.2). In this framework, emotional prosody is processed sequentially across time and neural structures. The first stage, occurring before 100 ms, involves early sensory analysis in

bilateral auditory cortices, where the LH exhibits higher temporal resolution and the RH greater spectral sensitivity. In the second stage (around 200 ms), emotionally salient acoustic cues—such as pitch and intensity—are integrated in the superior temporal gyrus (STG) and superior temporal sulcus (STS), forming an 'emotional gestalt'. This integration is thought to rely more heavily on RH structures, particularly for paralinguistic information, aligning with the acoustic asymmetry hypothesis. The third stage (beginning around 400 ms) recruits frontal regions, including the right inferior frontal gyrus (IFG) and orbitofrontal cortex (OFC), for evaluative judgments, while the left IFG supports semantic processing. This tripartite model highlights that prosodic processing is shaped by both temporal dynamics and cognitive context.

This aligns with the acoustic asymmetry hypothesis, where the RH preferentially processes slow, melodic contours (e.g., fundamental frequency, F0) and the LH is tuned to rapid, temporal features (Lancker & Sidtis, 1992; Zatorre, Belin, & Penhune, 2002). Zatorre, Evans, Meyer, and Gjedde (1992) showed RH prefrontal activation in response to pitch changes, supporting the RH's critical role in melodic analysis. In parallel, Gandour et al. (2004) proposed that lateralization also depends on linguistic experience and task demands, with corpus callosum-mediated interhemispheric transfer modulating LH involvement for complex auditory analysis.

A meta-analysis (Belyk & Brown, 2014b) further revealed that the right posterior superior temporal gyrus (pSTG) plays a central role in both emotional and linguistic prosody, although different prosodic functions may engage distinct neural networks. Temporal-lobe regions show stronger lateralization than frontal evaluative areas, suggesting a localizationist rather than strictly hemispheric view. Finally, Sammler, Grosbras, Anwander, Bestelmeyer, and Belin (2015) identified two prosody-processing pathways in the RH: a dorsal stream for temporally-sensitive acoustic cues ("how") and a ventral stream for integrating holistic prosodic patterns ("what"), paralleling the dual-stream model proposed for language processing in the LH (Hickok & Poeppel, 2004).

Taken together, these findings suggest that prosodic perception relies on dynamic and partially overlapping networks, with hemispheric specialization shaped by cue type, temporal structure, and task demands.

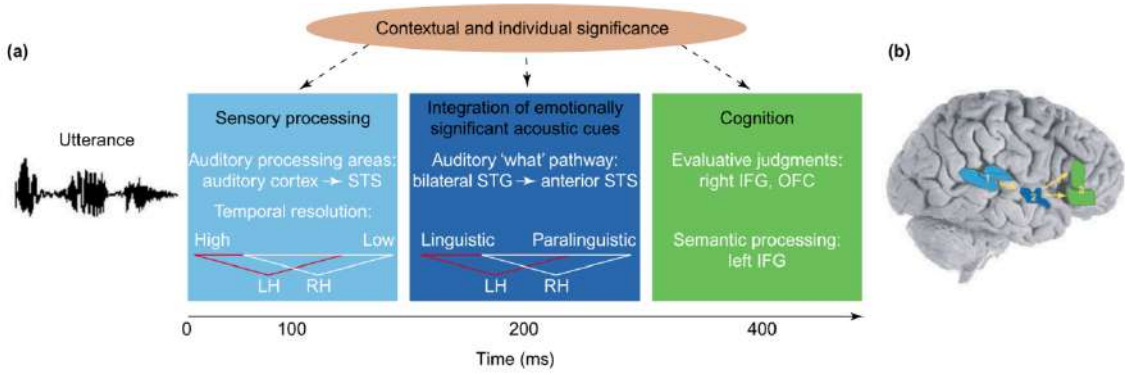


Fig. 2.2 **The three-stage model of emotional voice processing**, where the perception of prosody is processed along a right-hemisphere temporo-frontal gradient ranging from the superior temporal silcus (STS), superior temporal gyrus (STG) to the inferior frontal gyrius (IFG). Figure adapted from (Schirmer & Kotz, 2006)

2.2 Brain strokes: etiology, consequences and recovery

2.2.1 Etiology

Strokes, or cerebrovascular accidents, are the second leading cause of death and the primary cause of disability and loss of autonomy in France (Tuppin et al., 2016). In 2022 in France, 122000 adults were hospitalized for a stroke, and 30000 of them died (Lecoffre et al., 2017). While high blood pressure, obesity, physical inactivity and age are important risk factor, more than one in four stroke patients were aged < 65 years (Gabet et al., 2024).

Stroke results from a sudden disruption of cerebral blood flow, leading to insufficient oxygen and glucose delivery to brain tissue. Given the brain's high metabolic demand, this disruption can cause irreversible neuronal damage, known as infarction. Strokes are broadly classified into two main types. Ischemic strokes, which account for approximately 85% of cases, result from a reduction in blood flow to a specific region of the brain, typically due to the blockage of a cerebral artery. Hemorrhagic strokes, representing the remaining 15%, occur when a weakened blood vessel ruptures, often as a consequence of chronic hypertension or vascular abnormalities, leading to bleeding within or around the brain (Johns, 2014). In both types of stroke the brain tissue normally supplied by the vessel is suddenly deprived of blood and its function is lost, causing rapid-onset neurological

deficits (e.g., sudden weakness or loss of speech) (Kemmerer, 2022). These pathologies lead to structural, molecular, and functional alterations in brain tissue. The severity and range of neurological deficits that follow depend largely on the size and location of the brain lesion (Sperber, Gallucci, Mirman, Arnold, & Umarova, 2023).

2.2.2 Functional consequences

If a stroke damages critical brainstem regions essential for basic life functions, the individual may lose consciousness and die within minutes (Yuan et al., 2018). However, when the affected area is limited to cortical or white matter tracts responsible for specific cognitive abilities, the person typically survives but may experiences deficits in those functions. For instance, strokes involving the middle cerebral artery, which supplies essential language areas, typically result in aphasia (Kemmerer, 2022).

More generally, and depending on its severity and location, surviving a brain stroke is associated with a range of physical, cognitive/communication and emotional disabilities.

2.2.2.1 Motor impairments

Motor impairments, such as muscle weakness or hemiplegia, are common and can severely affect mobility and autonomy. Some patients may also display repetitive or rhythmic motor behaviors, leading to perseverative patterns of varying complexity (Li & Malhotra, 2015).

2.2.2.2 Cognitive impairments

Cognitive impairment affects nearly two-thirds of stroke survivors (Jin, Di Legge, Ostbye, Feightner, & Hachinski, 2006) and is particularly pronounced in attention, executive function, and processing speed—domains (Feigin et al., 2010) that are highly interdependent. A deficit in one, such as attentional control, can disrupt others. These challenges are notably evident in individuals with right hemisphere damage (RHD) with 65% impact (Stockbridge et al., 2022), who often present with executive dysfunction, sustained and selective attention deficits, unilateral neglect, and working memory impairments (McNab

& Klingberg, 2008). Approximately 96% of adults with RHD undergoing rehabilitation show deficits in at least one cognitive-communication domain (Blake et al., 2002; Tompkins, 2012), with such impairments having a marked impact on functional recovery and quality of life.

Memory impairments (Maeshima & Osawa, 2021), including working memory and recall deficits, are commonly observed in stroke patients and often co-occur with other cognitive-communication challenges. LHD patients may be primarily impaired in verbal memory (Blake et al., 2002). Conversely, (Gillespie, Bowen, & Foster, 2006) reveal that RHD patients show deficits in both verbal and non-verbal memory when compared to non-stroke controls, particularly on recognition tasks, which suggests difficulties in early encoding or storage of information.

Post-stroke *attention* deficits are highly prevalent, affecting up to 92% of individuals in the acute phase, and are associated with poorer motor recovery, increased fall risk, and reduced independence in daily activities (Barker-Collo, Feigin, Lawes, Senior, & Parag, 2010; Hyndman & Ashburn, 2003; Hyndman, Pickering, & Ashburn, 2008). Attention itself is a multifaceted construct comprising several interrelated subtypes. Focused attention allows individuals to respond to discrete stimuli, while sustained attention (or vigilance) enables prolonged engagement over time. Selective attention supports filtering out irrelevant input, and alternating attention enables cognitive flexibility by shifting between tasks or mental sets. Finally, divided attention, the most demanding form, involves managing multiple sources of information simultaneously (Cramer, Richards, Bernhardt, & Duncan, 2023). After a stroke, sustained and selective attention are particularly impacted, with 37% to 44% of patients showing impairments (Hyndman et al., 2008). These deficits appear to be especially common after right-hemisphere lesions, making it essential for researchers to ensure that RHD participants can meet the attentional demands of production and comprehension tasks (Heilman & Abell, 1980; Spaccavento et al., 2019).

Also related to attention and executive dysfunction, the symptom of *perseveration* is a frequent symptom following right-hemisphere stroke, especially in patients with co-occurring spatial neglect (Gandola et al., 2013; Nys, Zandvoort, Worp, Kappelle, & Haan, 2006). It manifests as inappropriate repetition of behavior or motor responses and can

significantly affect communication and daily functioning. Perseveration encompasses two primary types, recurrent markings (RM) and continuous markings (CM) (Sandson & Albert, 1984). RM perseveration reflects the delayed reactivation of a prior response, seen as multiple distinct strokes on the same target. CM perseveration refers to continued movement after task completion, appearing as excessive, uninterrupted strokes on a single item. Recent studies show that RM, but not CM, correlates with spatial neglect severity, suggesting RM may reflect difficulties in spatial disengagement, while CM involves failures in motor inhibition (Caulfield, Chen, Barry, & Barrett, 2017). Several established tasks exist for quantifying perseveration, such as Object Alternation (OA) (Freedman, Black, Ebert, & Binns, 1998) and the Wisconsin Card Sorting Test (WCST) (Abbruzzese, Ferri, & Scarone, 1996), which are widely used in assessing cognitive rigidity in conditions like aphasia and schizophrenia. These tasks measure executive function impairments, including difficulty in adapting to rule changes and excessive response repetition. As will be seen in Chapter 4, 8 and 9, perseveration will become a primary, albeit originally unexpected, object of study in this thesis.

2.2.2.3 Communication disorders

Communication and language disorders are common after stroke and significantly impact social participation and interpersonal relationships (Worrall et al., 2010).

Aphasia, primary associated with left-hemisphere strokes, affects language production, comprehension, reading, and writing, and can have lasting effects on autonomy and social interactions (Hamilton, Chrysikou, & Coslett, 2011). Aphasia is one of the most frequent and severe language disorders after stroke, affecting approximately 30˘34% of stroke patients in acute and rehabilitation settings, with frequencies as high as 62% in acute ischemic stroke when patients arrive within 3 hours of onset (Flowers et al., 2016). Over the long term, aphasia is associated with increased mortality, disability, and reduced likelihood of returning home. It results in impairments across language comprehension, production, and repetition, including difficulties retrieving words, forming grammatically correct sentences, or understanding spoken and written language. The traditional neurobiological model of aphasia is the Wernicke-Lichtheim model, which originated in the late

19th century as a method of synthesising neuropsychological findings of Broca, Wernike, and others (Eling, 2011), and is now generally considered to be linguistically and anatomically underspecified (Tremblay & Dick, 2016). This historical perspective is illustrated in Figure 2.3, which contrasts Wernicke's original diagram with Geschwind's influential 20th-century update emphasizing the arcuate fasciculus and the angular gyrus. Modern aphasiology research has revealed high variability within subtypes and limited correspondence between lesion sites and language profiles, challenging the validity of rigid diagnostic categories (Landrigan, Zhang, & Mirman, 2021). In particular, historically, the left hemisphere has been considered dominant for language processing, especially in right-handed individuals. However, accumulating evidence shows that the right hemisphere also plays a role in language, particularly in supporting pragmatic, prosodic, and contextual aspects of communication, and in aiding recovery after left hemisphere lesions (Bunker & Hillis, 2022).

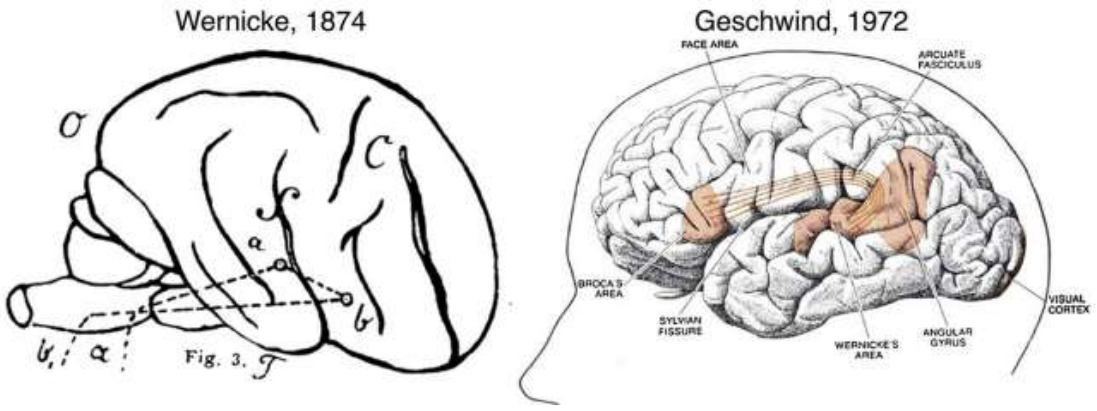


Fig. 2.3 **Language models from Wernicke to Geschwind Left:** Wernicke's original diagram **Right:** Geschwind's model, highlighting the arcuate fasciculus linking Broca's and Wernicke's areas, adapted from (O'Sullivan et al., 2019)

While much stroke research has centered on the left hemisphere, there is a growing acknowledgment of the communicative deficits that can arise from right hemisphere damage. Unlike aphasia which often manifests itself overtly, the challenges faced by individuals with right hemisphere damage are often less visible (Minga et al., 2023). These difficulties may only become evident through close interactions or observations by family members, who notice changes in their loved one's ability to engage in meaningful spoken discourse,

a monotone discourse and reduced participation in everyday communication.

Communication deficits associated with RHD include linguistic challenges such as maintaining discourse coherence, selecting appropriate words, constructing syntactic structures, and managing conversational topics (Davis, O'Neil-Pirozzi, & Coon, 1997). Impairments in the extralinguistic domain, such as difficulties in displaying appropriate emotional facial expressions and body language, further hinder effective communication. Paralinguistic deficits also play a significant role, with patients struggling to interpret and use non-literal language like idioms, metaphors, and sarcasm (Heath & Blonder, 2005). They may also have difficulty asking or understanding questions, exhibit tangential or egocentric discourse, and display verbosity or a paucity of speech.

Most notable among right-hemisphere communication deficits, and the key object of study in the present thesis, is the symptom of *aprosodia*, the inability to use or interprete speech prosody (Stockbridge et al., 2022; Ukaegbe et al., 2022). Aprosodia is estimated to affect 50%–78% of individuals with right hemisphere damage (Benton & Bryan, 1996; Cancelliere & Kertesz, 1990; Côté et al., 2007; Ukaegbe et al., 2022). It has been classified by Ross et al. (1997) into subtypes paralleling aphasia (e.g., motor, sensory, global as showed in Figure 2.4). Expressive aprosodia is often linked to right anterior lesions, while receptive forms are associated with right posterior damage. However, as for aphasia, these clinical categories often fail to reflect real-world complexity, leading researchers to adopt cognitive-neurofunctional models that emphasize hierarchical auditory processing and distributed neural networks (Baum & Pell, 1999). As seen in section 2.3 below, the clinical assessment of aprosodia is a bit of a weak-point in post-stroke rehabilitation, with several existing tools that are plagued by important methodological limitations.

Somehow related to aprosodia, stroke survivors also have been described to suffer from acquired *amusia*, of the inability to perceive and enjoy music. Music perception share overlapping neural mechanisms with prosodic processing, especially in the right hemisphere. Both amusia and aprosodia have been associated with damage to the right frontoinsular cortex, striatal regions, and disconnection of the right inferior fronto-occipital fasciculus (Hausen, Torppa, Salmela, Vainio, & Särkämö, 2013; Sihvonen et al., 2021). These areas are part of the right ventral auditory stream, which integrates melodic and rhythmic

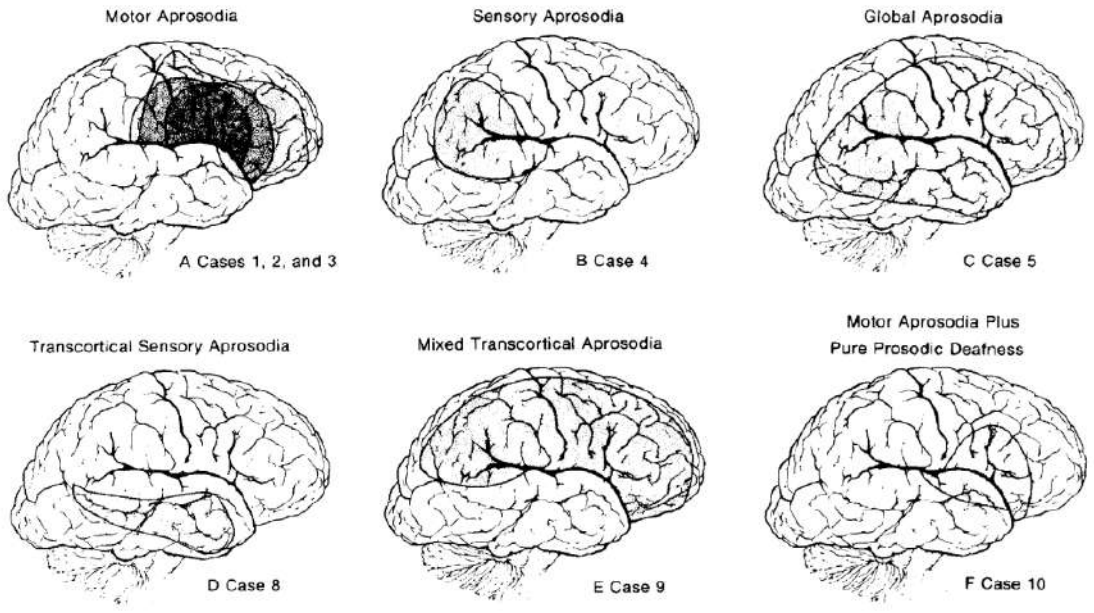


Fig. 2.4 **Right hemisphere lesion sites associated with different subtypes of aprosodia** adapted from (Ross, 1981)

information essential for interpreting both music and speech prosody. This overlap supports the view that pitch perception acts as a common foundation for both domains. The Montreal Battery of Evaluation of Amusia (MBEA) (Peretz, Champod, & Hyde, 2003), which includes subtests like contour and interval discrimination, is widely used to assess amusia and can offer indirect insights into prosodic impairments as well.

2.2.2.4 Psychiatric and emotional disorders

Finally, neuropsychiatric disorders, particularly *post-stroke depression* (PSD), are frequent after stroke and have a profound impact on quality of life (Oliveira et al., 2015). A pilot study reported that 22.5% of stroke patients developed PSD within the first three months following a stroke. The basal ganglia, which modulate both mood and speech motor control via dopaminergic pathways, may explain the frequent co-occurrence of mood disorders and aprosodia, especially in expressive and receptive deficits seen in post-stroke depression (Uekermann, Abdel-Hamid, Lehmkämper, Vollmoeller, & Daum, 2008). Acoustic analysis of the voice of stroke survivors with post-stroke depression revealed reduced variability

in fundamental frequency along with significant alterations in voice breaks and shimmer, both of which were strong predictors of PSD risk. Early changes in affective prosody have been closely linked to an increased likelihood of developing PSD within the first period after stroke (Villain et al., 2016). These mood symptoms are often evaluated using the Hospital Anxiety and Depression Scale (HADS), for detecting symptoms of anxiety and depression (S. Friedman, Samuelian, Lancrenon, Even, & Chiarelli, 2001). Finally, on top of all previous symptoms, stroke survivors may also experience *anosognosia*, a meta-cognitive condition in which individuals are unaware of their own sensory, motor, or cognitive deficits (Vuilleumier, 2004).

2.2.3 Recovery after stroke

Rehabilitation after a stroke aims to address the cognitive, motor, and emotional consequences of the injury, and is most effective when delivered through multidisciplinary and multidomain programs tailored to the individual's deficits and recovery stage, with the goal of restoring autonomy and improving quality of life (Benedetti et al., 2022; Langhorne, Bernhardt, & Kwakkel, 2011; Licht, 1975).

Physiotherapy and targeted exercises are often employed to restore motor function (Ernst, 1990). Speech and language therapy plays a central role in managing cognitive and communication disorders. It targets impairments in receptive and expressive language (understanding and producing spoken language), as well as reading, writing, and functional communication (everyday use of language in context) (Brady, Godwin, Enderby, Kelly, & Campbell, 2016). Finally, cognitive and behavioral therapy (CBT) is often used to support recovery from post-stroke depression, anxiety, and fatigue, which may co-occur with cognitive symptoms (Cumming, Packer, Kramer, & English, 2016).

In addition to traditional therapies, neuromodulation techniques such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) have shown promise in enhancing recovery and are now recommended in chronic-phase stroke rehabilitation by the French health authority (HAS) (Saway et al., 2024).

2.3 The clinical assessment of aprosodia: aim of the thesis

This thesis focuses on the specific post-stroke symptom of aprosodia, and how to improve its clinical assessment by introducing a new experimental procedure based on the reverse-correlation paradigm.

Because of their low visibility compared to the more manifest disorders of aphasia, comprehensive diagnostic tools are crucial for detecting and managing aprosodia effectively. Speech-language pathologists are trained to assess and treat cognitive-communication disorders, however patients are often not referred due to the subtle and easily overlooked nature of these deficits in clinical settings (Blake et al., 2002).

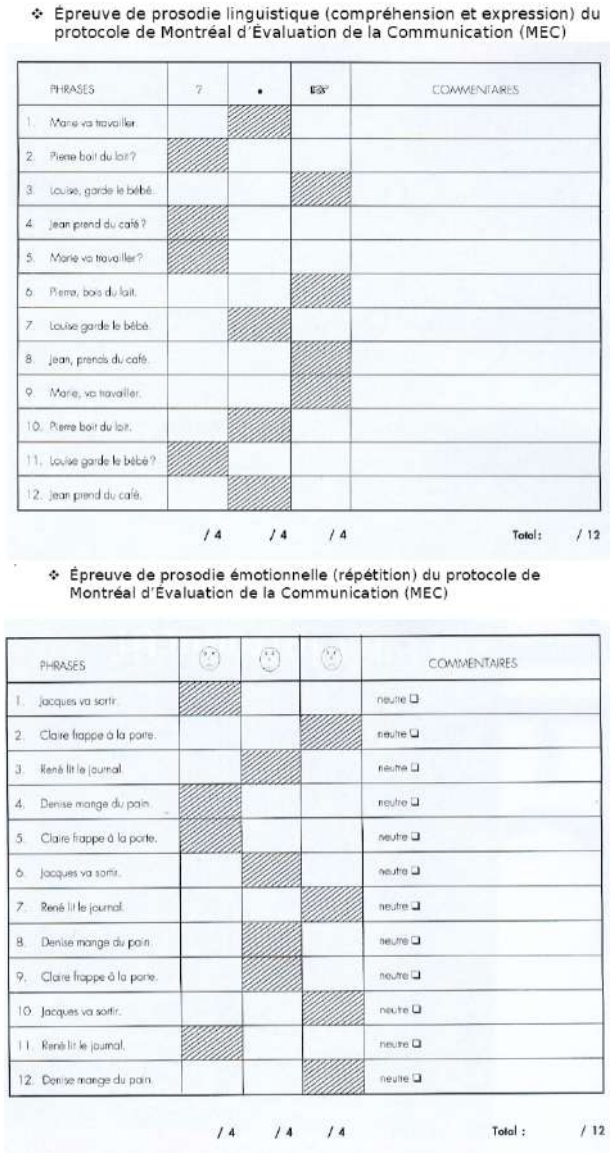
2.3.1 Existing tools for the assessment of aprosodia

Several batteries exist to evaluate prosodic deficits, though they remain limited in availability, especially across languages (Benedetti et al., 2022). Key tools include the Aprosodia Battery (Ross et al., 1997), the Battery of Emotional Expression and Comprehension-BEEC- (Cancelliere & Kertesz, 1990), the New York Emotion Battery-NYEB- (Borod, Welkowitz, & Obler, 1992), the Montreal Evaluation of Communication-MEC- (Joanette et al., 2004), the Assessment Battery for Communication-ABaCo- (Angeleri et al., 2008), the Affective Communication Test-ACT- (H. S. Friedman et al., 1980), the Florida Affect Battery-FAB- (Bowers, Blonder, & Heilman, 1998), and the Lille Communication Test-LCT- (Rousseaux, Daveluy, & Kozlowski, 2010).

Most of these assessment tools rely on neuropsychological tasks (e.g., "paper-and-pencil" measures) to evaluate prosodic abilities. However, these tools suffer from several methodological limitations. Manual scoring introduces inter-rater variability, and many assessments lack the sensitivity and reproducibility required to track longitudinal changes or therapy effectiveness reliably (Benedetti et al., 2022).

Second, most batteries primarily assess emotional prosody, while linguistic prosody remains underrepresented, appearing only as subtests in FAB, MEC, ABaCo, NYEB, and LCT. A few tools (e.g., BEEC, FAB, ACT, NYEB) also evaluate facial emotion perception and expression. Expressive prosody tasks often rely on subjective clinician ratings of

facial expression or vocal tone, while receptive prosody assessments are challenging due to the complexity of decoding emotional intent. Notably, only three tools, the BEEC, Aprosodia Battery, and NYEB, specifically emphasize receptive prosody (Benedetti et al., 2022). Finally, most of these tools are published in English; only a few are available



❖ Épreuve de prosodie linguistique (compréhension et expression) du protocole de Montréal d'Évaluation de la Communication (MEC)

PHRASES	?	.	!	COMMENTAIRES
1. Marie va travailler.				
2. Pierre boit du lait?				
3. Louise, garde le bébé.				
4. Jean prend du café?				
5. Marie va travailler?				
6. Pierre, bois du lait.				
7. Louise garde le bébé.				
8. Jean, prends du café.				
9. Marie, va travailler.				
10. Pierre boit du lait.				
11. Louise garde le bébé?				
12. Jean prend du café.				
	/ 4	/ 4	/ 4	Total: / 12

❖ Épreuve de prosodie émotionnelle (répétition) du protocole de Montréal d'Évaluation de la Communication (MEC)

PHRASES				COMMENTAIRES
1. Jacques va sortir.				neutre ☐
2. Claire frappe à la porte.				neutre ☐
3. René lit le journal.				neutre ☐
4. Denise mange du pain.				neutre ☐
5. Claire frappe à la porte.				neutre ☐
6. Jacques va sortir.				neutre ☐
7. René lit le journal.				neutre ☐
8. Denise mange du pain.				neutre ☐
9. Claire frappe à la porte.				neutre ☐
10. Jacques va sortir.				neutre ☐
11. René lit le journal.				neutre ☐
12. Denise mange du pain.				neutre ☐
	/ 4	/ 4	/ 4	Total : / 12

Fig. 2.5 **Examples of two prosody-related subtests from the Montreal Evaluation of Communication (MEC) protocol. Top:** Comprehension task assessing the patient's ability to interpret prosodic cues in spoken sentences. **Bottom:** Repetition task requiring patients to reproduce the prosody of a model sentence. In this case, both tasks were scored at 12/12, indicating full performance. However, these tasks primarily test recognition and imitation, with limited assessment of spontaneous or internally generated prosodic contours.

in French (e.g., MEC, LCT) or other languages such as Italian (ABaCo), limiting their international applicability.

The Montreal Evaluation of Communication (MEC) protocol (Joanette et al., 2004) is currently the reference tool for evaluating communication deficits in French-speaking adults following right hemisphere stroke. While it includes both listening and production tasks, such as repetition and comprehension subtests (Figure 2.5), its prosody section presents notable shortcomings. Specifically, it evaluates repetition of pre-recorded prosodic contours but fails to assess the patient's ability to perceive or generate prosody independently of acoustic targets. As a result, it often produces false negatives and lacks diagnostic precision for prosodic deficits (Rosenbek et al., 2004). Furthermore, inter-rater agreement, although generally good across the MEC, is lower for prosodic components (Côté et al., 2007).

Several alternative tools exist, such as the Battery of Emotional Expression and Comprehension (BEEC; Cancelliere & Kertesz, 1990) and the Aprosodia Battery (Ross et al., 1997), but these focus exclusively on emotional prosody. The Florida Affect Battery (Bowers et al., 1998) includes a linguistic prosody section but has not been validated in French, limiting its clinical applicability in francophone contexts.

Additionally, most available tools provide only categorical outputs (presence/absence of impairment), offering limited insight into the underlying cognitive or perceptual mechanisms. This lack of granularity hinders the development of targeted rehabilitation strategies and limits their use in research contexts where mechanistic understanding is crucial.

2.3.2 Aim of the thesis

This thesis aims to develop a more sensitive method for evaluating prosodic perception deficits, with a particular focus on the linguistic aspects of prosody, which have been comparatively underexplored in the context of right hemisphere stroke. Existing tools often lack the precision and adaptability required to capture subtle impairments in this domain, particularly in clinical settings.

To address these limitations, we propose a novel experimental procedure that avoids

traditional repetition or imitation tasks and instead focuses on patients' spontaneous perceptual judgments. Our approach is based on probing participants' internal perceptual representations, using the psychophysical methodology of *reverse correlation* (see Chapter 3, allowing for a more ecological and cognitively efficient assessment that aligns more closely with real-world communication demands.

Importantly, this approach is not intended to replace existing clinical batteries but to complement them by offering mechanistic insights into the underlying sensory and cognitive processes of prosodic perception. By combining this novel perspective with traditional tools, clinicians and researchers may gain a more comprehensive understanding of individual deficits.

Ultimately, this work addresses a key gap in current assessment practices and strives to contribute a robust framework for identifying and characterizing prosodic deficits in neurological populations.

CHAPTER 3

Methodological foundations: The reverse-correlation method

3.1 Psychophysics, signal-detection theory and the computational modeling of sensory decisions

3.1.1 Fechner, Weber and Stevens: a very short history of psychophysics

The mind and body, though different in nature, are not separate entities but can in theory be linked by mathematical relations. This - a form of dual-aspect monism (Crane & Patterson, 2012)- was the belief of 19th-century German physicist/psychologist Gustave Fechner, who sought to uncover the relationship between the physical nature of sensory neurons and the mental processes that give rise to thought, emotions or consciousness. Fechner strongly believed in the power of empirical evidence, and his ambition was to find a way to measure the mind. In 1860, he laid the foundation for the field psychophysics

(Fechner, 1948), a science dedicated to understanding the relationship between physical stimuli and mental phenomena. Psychophysics not only investigates how our sensory organs operate but also explores how the mind interprets sensory information.

Quantitative measurements started very early in the 1830s with Fechner conducting detection tasks for the presence or absence of a stimulus using the concept of the absolute threshold, which defined the minimum intensity of a stimulus detectable 50% of the time based on the decision criterion of one (Fechner, 1948). This was complemented by Ernest Weber's introduction of discrimination tasks, particularly the difference threshold and the just-noticeable difference (JND). These tasks, often conducted using formats as two-alternative forced choice (2AFC, i.e., one presented stimulus and two response options: participant has to choose whether the stimulus contains the target stimulus) or two-interval forced choice (2IFC, i.e., two stimuli presented sequentially, requiring the participant to identify the one containing the target), quantified the smallest detectable change in a stimulus to find the participant's detection threshold.

One of the first proposed principles of psychophysics is Weber's Law, which established that the JND (ΔI) is proportional (k) to the stimulus intensity (I) (Falmagne, 1985).

$$\frac{\Delta I}{I} = k \tag{3.1}$$

In 1860, Fechner extended this work with Fechner's Law (Johnson, Hsiao, & Yoshioka, 2002), proposing that perceived intensity (S) scales logarithmically with physical stimulus intensity (I), linking perception to cumulative JNDs.

$$S = k \log(I) + C \tag{3.2}$$

The results of such experiments are often visualized using a psychometric curve, which plots the probability of a participant's response (e.g., detection or discrimination) against the physical intensity of the stimulus. This curve, typically S-shaped, allows researchers to estimate thresholds (e.g., the absolute threshold or JND) and analyze how sensitivity changes with stimulus intensity.

Later, in 1957, Stevens' Power Law (Stevens, 1957) generalized Fechner's logarithmic model, showing that perception could follow linear, logarithmic, or exponential relation-

ships depending on the sensory modality. Stevens employed scaling tasks where participants estimated perceived intensities using rating scales or proportional judgements.

$$S = kI^n \tag{3.3}$$

These milestones, integrating detection, discrimination, and scaling tasks, formed the foundation of modern psychophysics, advancing our understanding of sensory thresholds, perceptual scaling, and the intricate relationship between physical stimuli and subjective experience.

3.1.2 Signal detection theory

The development of Signal-detection theory (SDT) in 1966 by Green and Swets (1966) marked a significant advance in psychophysics. SDT addressed limitations in earlier methods, which could not effectively estimate false positives (false alarms) or distinguish them from true detections. Traditional threshold-based approaches indeed assumed that detection was solely based on sensory sensitivity and failed to account for decision-making factors.

Unlike traditional threshold-based approaches, SDT introduced a probabilistic framework, emphasizing that detection depends on both the actual sensory signal and the subject's decision-making process. Key tools, like Receiver Operating Characteristic (ROC) curves (Hanley et al., 1989), allowed researchers to analyze the trade-off between hit rates ($z(H)$, where z is the so-called z-transform, i.e., the inverse of the standard normal cumulative distribution) and false alarms ($z(FA)$), separating perceptual sensitivity (d-prime, or d') and response bias β.

$$d' = z(H) - z(FA) \tag{3.4}$$

$$\beta = -\frac{z(H) + z(FA)}{2} \tag{3.5}$$

3.1.3 The linear observer model

Signal Detection Theory (SDT) is based on how we detect a signal from various sensory stimuli amidst noise ($signal + noise$), and is concerned with the correct measurement

of sensory performance and separating it from stimulus-independent decision-making parameters such as response bias. However, SDT does not provide a computational model of observer decisions, i.e., does not describe what exact stimulus features are used to make a sensory decision and how these features are weighted by the observer's sensory/perceptive system. Developing such observer models has been a major concern of the modern field of psychophysics (Lu & Dosher, 2008).

One simple but prominent observer model in psychophysics in the linear observer model (Abbey, Eckstein, & Bochud, 1999; Burgess, Wagner, Jennings, & Barlow, 1981). In this model, an observer is assumed to have internal templates (arguably, 'mental/sensory representations') representing the signals being presented and to make decisions by computing the similarity of the stimulus (input) to these templates. In other words, the observer's responses are a weighted linear combination of the stimulus features with added noise representing internal variability as we'll see below.

In more details, consider a 2IFC signal-in-noise detection experiment where two randomized signals s_1^t and s_2^t are presented in each trial t, and an observer is tasked to identify which of s_1,s_2 best matches an internal template k (alternatively called a kernel, see Section 3.2.1). A minimal model for how such an observer may come to a decision is to compute decision variables d_1^t and d_2^t by taking the dot product of the stimuli with the kernel $s \cdot k$, and adding realizations from an independent source of 'internal noise' n (called 'internal' in contrast to 'external noise', which is the noise applied experimentally to the stimuli s_1 and s_2):

$$d_1^t = s_1^t \cdot k + n_1^t \tag{3.6}$$

$$d_2^t = s_2^t \cdot k + n_2^t \tag{3.7}$$

The model assumes that the observer identifies the signal as s_2 if d_2 plus some constant b (response bias) is larger than d_1:

$$(s_2^t - s_1^t) \cdot k + (n_2^t - n_1^t) > b \tag{3.8}$$

or, equivalently,

$$r^t = \begin{cases} 2 & \text{if } (s_2^t - s_1^t) \cdot k + n^t > b \\ 1 & \text{otherwise} \end{cases} \tag{3.9}$$

where there n is a source of internal noise with a variance σ_n that is twice the variance of n_1 and n_2.

Figure 3.1 provides an illustration of how the linear observer model can be used to generate/simulate observer responses: an observer with a given kernel (top) responds to 150 successive 2IFC trials (middle). For each trial, two stimuli are compared to the kernel, and the stimulus that best matches it (with additive internal noise) is chosen as the response.

Operating under the assumption of such a model, the task of psychophysical research therefore consists of obtaining estimates for observer parameters (template k and internal noise σ_n) given experimental measurements as well as evaluate the extent to which such a model fits actual observations and, if needed, explore more complex alternative models (for a review of alternative classical formulations of observer models, see (Lu & Dosher, 2008)). One common experimental approach to estimate these parameters is the so-called *reverse correlation* procedure.

Palin Toolbox: Simulating linear observers

In this chapter and the following, we will illustrate and analyse reverse-correlation data using the methodology of computer simulation and a specially developed Python toolbox (PALIN) which we developped for this purpose. PALIN introduces an object-oriented architecture for defining simulated observers (with their template, noise and bias parameters) and letting them encounter simulated experiments with random trials. The following code illustrates how Figure 3.1 was generated: we create a LinearObserver `obs` with a kernel, internal noise and bias; we then let the observer respond to a simulated 2IFC experiment (here, a `DoublePassExperiment`, see Section 3.2.3.1 below). This generates a list of responses (e.g., 0,1,0,0,0,1,1), corresponding to stimulus choice for each trial. An illustrative output of the simulation is given in Figure 3.1.

```
# Define a DoublePassExperiment with 100 trials + 50 repeated trials
exp = DoublePassExperiment(n_trials = 100, n_repeated = 50,
                           trial_type = Int2Trial,
                           n_features = 6,
                           external_noise_std = 100)
# Create an Observer with a random kernel and internal noise
obs = LinearObserver.with_random_kernel(n_features = exp.n_features,
                                        internal_noise_std = 3,
                                        criteria = 0)
# Generate responses
responses = obs.respond_to_experiment(exp)
# Convert experiment and responses to a DataFrame for analysis
responses_df = Analyser.to_df(exp, responses)
```

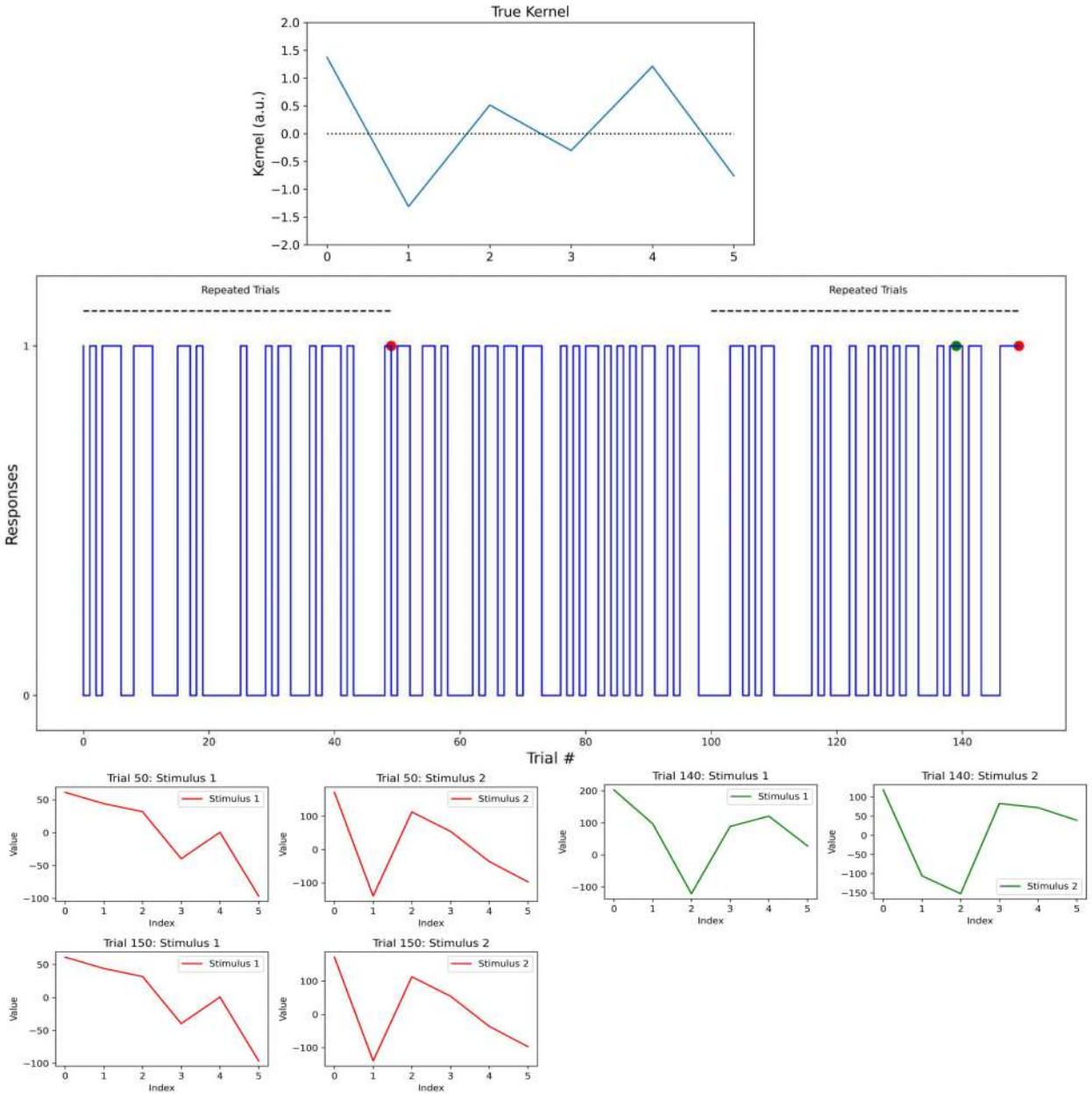


Fig. 3.1 **Simulation of a Reverse-correlation experiment with Double-pass blocks Top:** The PALIN toolbox can simulate reverse-correlation experiments with various stimulus presentations, such as 1AFC and 2AFC tasks. A linear observer with a random kernel is used to generate responses to the stimuli. **Middle:** Illustration of responses of a linear observer with its defined kernel to the simulated experiment. These simulations can include repeated blocks of trials (double-pass) to assess response consistency. In this example, trials 0–50 (first block) are repeated in trials 100–150 (third block). **Bottom:** Illustration of stimulus generation in a 2AFC reverse-correlation experiment. Trials 50 and 150 share identical stimuli, although the simulated observer responds differently (response = 0 at trial 50, response = 1 at trial 150). In contrast, a different stimulus at trial 51 yields the same response as trial 150, suggesting inconsistency in the observer's responses.

3.2 Reverse correlation

Reverse correlation is an experimental approach that aims to estimate an observer's decision parameters in a data-driven manner. When operating under the assumption of the linear observer model, reverse correlation provides both an experimental procedure, as well as analytical methods, to estimate both kernel k, internal noise σ_n and bias b from a series of stimulus-response pairs $(s_1^t, s_2^t; r^t)$ corresponding to how a given observer responded to a given experiment.

3.2.1 The white-noise method, and Wiener/Volterra kernel theory

The idea of reverse correlation actually comes from the field of neurophysiology (Ringach & Shapley, 2004). Originally called the *white-noise method*, it was used to measure the receptive fields of visual neurons by analyzing their responses to random inputs, such as white noise. This approach has provided critical insights into spatial and color processing in the cortex at low-level processing in early sensory areas as V1 (Marmarelis & Marmarelis, 2011). In its neurophysiological origins, random stimuli are presented to a sensory neuron, and sparsely occurring neuronal spikes are measured in output **??**. By averaging the (random) stimuli that generated a response, the spike-triggered average (STA) is obtained, representing the weights of a linear filter that approximates the neuron's response and reflects its preferred stimulus under the assumption of linearity - essentially, as we'll see below, a template k in the sense of the linear observer model.

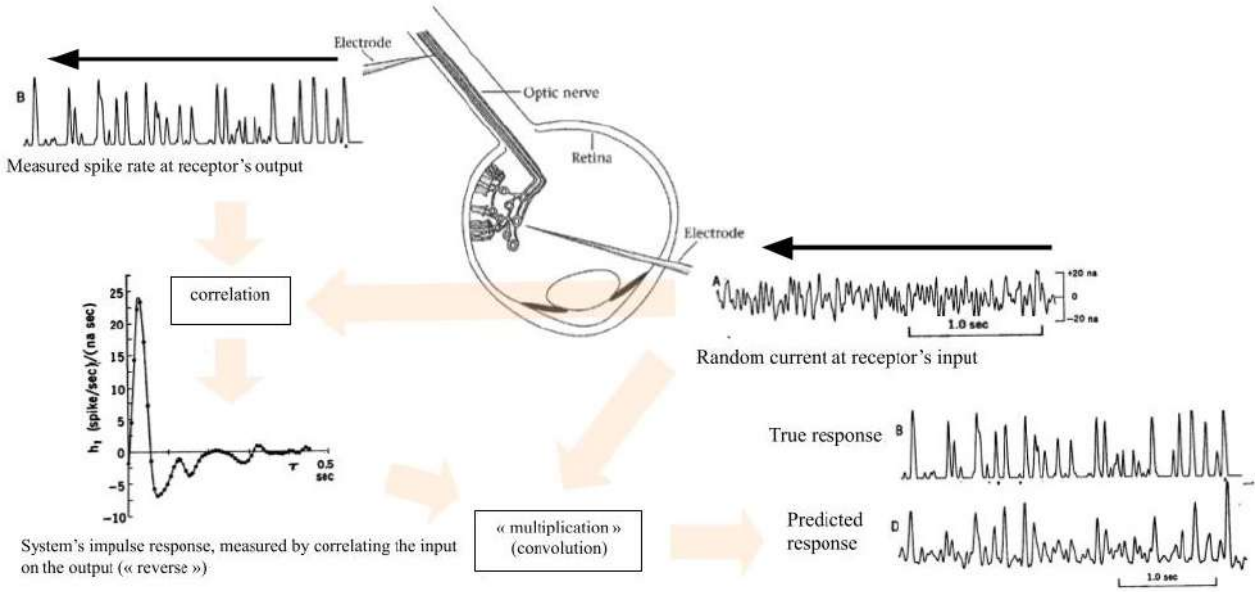


Fig. 3.2 **Reverse correlation in the early visual system** This diagram illustrates how a neuron's stimulus–response properties can be characterized using system identification techniques. A random white noise current is injected at the receptor level of the retina (top right), and the resulting spike activity is recorded from the optic nerve (top left). The measured spike rate reflects the output of the neuron in response to the dynamic input. By computing the cross-correlation between the stimulus and the neural response, we estimate the system's linear temporal filter, also called the impulse response or kernel (bottom left). This kernel captures how the input is temporally integrated to generate the output. Once this kernel is estimated, it can be used to predict the system's response to new stimuli through convolution (bottom right). This approach, often referred to as the reverse correlation method.

An alternative mathematical view of the white-noise method is the Wiener/Volterra theory of system identification (Eggermont, 1993). Volterra (1930) and Wiener (1966) demonstrated that under certain conditions (e.g., time invariance and finite memory), complex systems with time-varying inputs $x(t)$ and outputs $y(t)$ can be approximated as a sum of simpler subsystems. These include a zero-order subsystem, which produces a constant output; a first-order subsystem, which generates a weighted sum of past inputs using what is called a first-order kernel; a second-order subsystem, which captures pairwise interactions of past inputs through the second-order kernel, and so forth. Lee and Schetzen (1965) showed that white noise input, having a flat power spectral density, enables the direct evaluation of kernels by correlating input and output. For instance, the zero-order kernel corresponds to the average output, the first-order kernel is derived from the correlation between input and output, and the second-order kernel involves correlations between pairs of inputs and the residual output. Here again, the rational is that one can estimate a first-order linear approximation of a system - e.g., an experimental observer - by presenting them with random stimulus variations and analyzing the set of their responses. This provides a way to estimate a linear-observer template k as the first-order Volterra/wiener kernel of the corresponding stimulus-response system. In the psychophysical literature as well as the rest of this thesis, the word *kernel* is actually alternatively used to describe an observer's template - hence, also the use of letter k in our formulation.

3.2.2 Estimating templates

Building on the white noise approach in neurophysiology, psychophysicists have proposed to estimate linear-observer templates/kernels by presenting them a large number of randomly manipulated stimuli, and by "reverse" correlating observer responses on stimulus characteristics. In a typical experiment, the stimulus consists of one of two possible signals embedded in a Gaussian noise field that varies from trial to trial. The observer's task is to identify which signal was shown on each trial, i.e., to provide a single discrete response (rather than a continuous variable output over time, as in neurophysiology).

3.2.2.1 The classification-image method

A simple method to estimate templates in such a setup is the so-called *classification image* method[1] - essentially a weighted sum. The process involves averaging the noise from trials with positive responses, and subtracting the average noise from trials with negative responses 3.10. In the case of graded responses (i.e., stimulus intensity instead of stimulus detection), the observer's response on trial t is modeled as a linear combination of the stimulus features, weighted by sensory weights. The weighted sum is replaced by a least-squares regression to relate observers' rating responses to the stimulus properties. The resulting regression coefficients can be interpreted as the sensory weights that observers assigned to different properties (e.g., frequency bins in a spectrogram stimulus - A. Ahumada and Lovell (1971)) when judging the presence of the signal.

$$\hat{r}^t = \sum_{i=1}^{d} s_i^t \cdot k_i + \epsilon^t \tag{3.10}$$

3.2.2.2 The Generalized Linear Model method

An equivalent view of Eq. 3.9 is to consider responses r_t as the binary outcome of a logistic regression (or, equivalently, a Generalized linear model - GLM), given by:

$$y^t = g(\beta_0 + \sum_{i=1}^{N} \beta_i x_i^t) \tag{3.11}$$

where the weights $\beta_{i=1...N}$ of the linear predictor correspond to the coordinates of kernel k, and input x_i correspond to stimulus data (in the 2AFC case of Eq. 3.9, x_i^t is the i^{th} coordinate of stimulus difference $s_1^t - s_2^t$), both of dimension N; and g a non-linear link function (logit or probit (Müller, 2011)).

Generalized Linear Models (GLMs) are a broad class of regression-like models that extend traditional linear regression to accommodate dependent variables with distributions

[1]The terminology in the field of reverse correlation suffers from some ambiguity. Some authors use the word "classification image" to refer to the estimated template, with the rationale that it provides a representation of how a (e.g., visual) classification is done, if such is the task presented to the observer (by extension, templates extracted from auditory experiments are also called auditory classification *image* (Varnet, Wang, Peter, Meunier, & Hoen, 2015)). Other authors use the "classification image" method to describe the averaging procedure by which the template is estimated (Murray, 2011). We use it here in this sense and will refer to the template as template or kernel.

beyond the normal, such as Bernoulli, binomial, or Poisson (McCullagh, 2019). The GLM framework models the expected value of a dependent variable as a nonlinear function of a linear combination of predictors through a link function, making it highly flexible and adaptable to a variety of data types.

One of the key strengths of GLMs lies in their reliance on Maximum-Likelihood-Estimation (MLE) for parameter fitting. This robust estimation technique determines the regression coefficients (β) by maximizing the likelihood of the observed data under the assumed model. The likelihood function is given by:

$$L(\beta \mid y, X) = \prod_{i=1}^{n} f(y_i \mid \mu_i), \tag{3.12}$$

where β is the vector of regression coefficients (including the intercept β_0), $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ is the vector of observed responses, and $\mathbf{X}$ is the matrix of input features or covariates. Each y_i corresponds to the observed response on trial i, and μ_i is the expected value of y_i under the model, typically defined as $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$, where g^{-1} is the inverse of the link function (e.g., logistic or probit). The function $f(y_i \mid \mu_i)$ denotes the likelihood of observing y_i given the model prediction μ_i, based on the assumed distribution of the GLM (e.g., Bernoulli for binary data). The product is taken over all n trials in the dataset.

The maximum likelihood estimates of β are obtained by solving:

$$\hat{\beta} = \arg\max_{\beta} \log L(\beta \mid y, X) \tag{3.13}$$

It indicates that we search for the value of β that maximizes the log-likelihood of the observed data $\mathbf{y}$ given the inputs $\mathbf{X}$ under the GLM model. Taking the logarithm of the likelihood simplifies the product in Equation 3.12 into a sum, which is more convenient for numerical optimization.

These estimates enable hypothesis testing and model comparison using statistical tools such as the Akaike Information Criterion (AIC), (Knoblauch & Maloney, 2008) which evaluates the relative quality of different models.

Knoblauch and Maloney (2008) explored the relationship between GLMs and visual classification images, demonstrating through simulations that GLMs offer a principled and robust framework for estimating observer templates. They found that GLMs are more

resistant to noise compared to traditional weighted sum methods. Similarly, Mineault, Barthelmé, and Pack (2009) extended these ideas to analyze complex decision-making tasks by including non-linear relationship wih sparse priors, enhancing GLM utility for sparse data and showing that GLMs can infer a psychophysical observer's decision process with fewer trials than previously proposed methods. This enables researchers to explore more sophisticated and informative models of decision-making processes (Okazawa, Sha, Purcell, & Kiani, 2018) while maintaining statistical tractability.

Palin Toolbox: Kernel estimation with classification images and GLM

PALIN can implement both the weighted sum (ClassificationImage) method and the GLM method for kernel estimation. The following code configures and runs a simulation for comparing the accuracy of both methods for experiments with 150, 500, and 1000 trials. It defines parameters for observers, experiments, and analyzers, and performs the simulation over all possible configurations. Stimuli are 6-dimensional, randomly sampled by a normal distribution with standard deviation of 100. Simulated responses are analysed to look at the error of estimation between 2 methods of kernel extraction and the true (known) observer kernel.

```
# Simulate multiple linear observer to look at the Correlation and confidence interval of estimation
observer_params = {'kernel':['random'],'internal_noise_std':np.arange(0,5.1,0.5), 'criteria':[0]}
experiment_params = {'n_trials': [150,500,1000], 'trial_type': [Int2Trial],
    'n_features': [6],'external_noise_std': [100]}
analyser_params = {'kernel_extractor': [ClassificationImage,GLMKernel],'distance': ['CORR']}
sim_kernel = Sim(SimpleExperiment, experiment_params,  LinearObserver, observer_params,
                KernelDistance, analyser_params)
sim_kernel_df = sim_kernel.run_all(n_runs=10)
```

The accuracy of kernel estimation using reverse correlation methods naturally depends on the number of trials available. Figure 3.3 compares the performance of two common approaches for kernel extraction: the weighted sum method or the classification image and the Generalized Linear Model (GLM). The y-axis shows the correlation between the estimated kernel and the true underlying kernel used to generate the data, while the x-axis represents the number of trials.

As expected, both methods improve in accuracy with increasing trial numbers. How-

ever, the GLM consistently outperforms the weighted sum method, particularly when the number of trials is limited (e.g., below 500). This advantage likely stems from the fact that the GLM explicitly models the probabilistic link between input and response, which leads to more robust estimates in low-data regimes. When the number of trials reaches 1000, the difference between the methods narrows, but the GLM still shows slightly higher correlation with the ground truth.

3.2.3 Estimating internal noise

Within the linear observer model, internal noise σ_n refers to the inherent variability (Neri, 2010) in the responses of the observer's sensory and decision-making system. This variability arises from the stochastic nature of neural activity, which limits the reliability and accuracy of perceptual systems. Interestingly, internal noise is not unique to biological systems, signal degradation due to variability is a well documented phenomenon in electronic systems, such as amplifiers. However, in sensory neurons, internal noise plays a critical role as a limiting factor in signal transduction, influencing both perception and behavioral performance (Faisal, Selen, & Wolpert, 2008).

Internal noise, distinct from external variability, is a critical factor in signal detection. When a signal is embedded in external noise (e.g., added Gaussian noise), an ideal observer uses an internal template to match the signal. In the absence of internal noise, this template closely aligns with the signal, enabling accurate detection. However, internal noise necessitates adjustments, such as clipping or rescaling the template, to maintain functionality. These adjustments illustrate how internal variability constrains the observer's ability to optimize detection, with higher internal noise limiting both the range and accuracy of the template (Neri, 2020).

The internal-to-external noise ratio provides insight into the relative contributions of sensory variability and environmental factors, indicating whether performance is more influenced by internal processes or external conditions. Understanding this relationship is critical for characterizing perceptual sensitivity and decision-making.

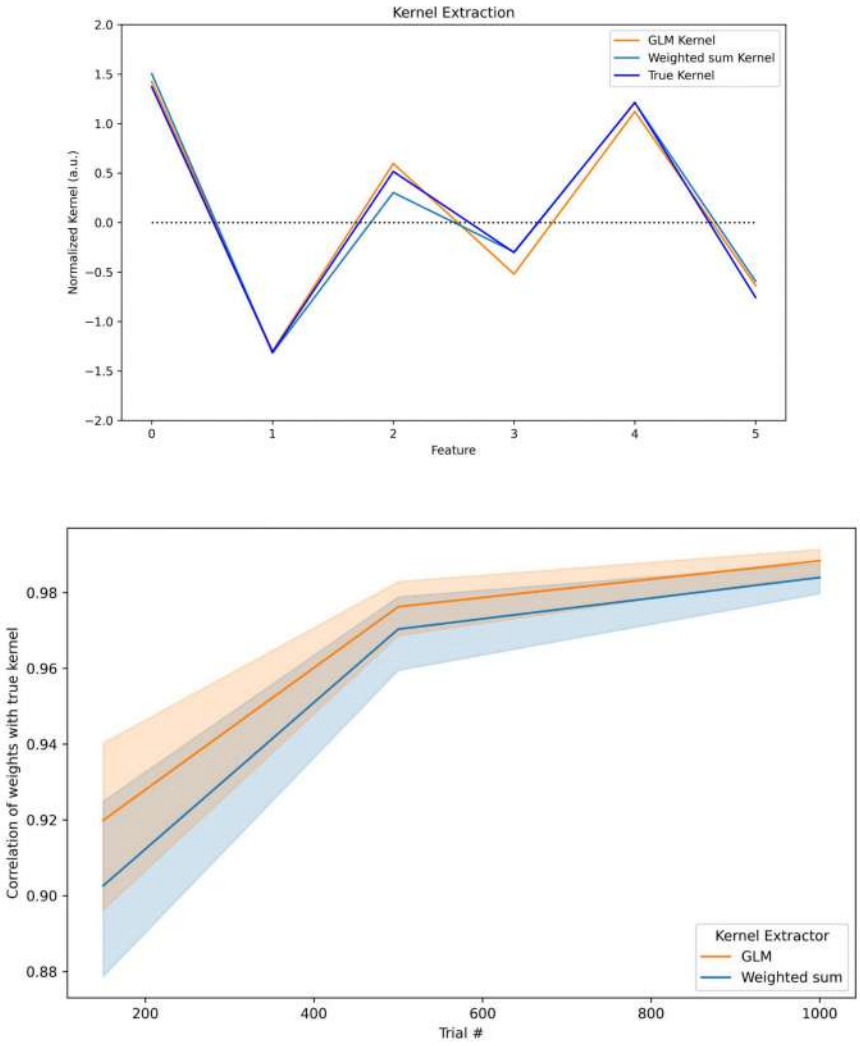


Fig. 3.3 **Comparison of kernel estimation methods and their sensitivity to the number of trials.Top:** Example of estimated kernels obtained with two methods of weighted sum (classification image) and Generalized Linear Model (GLM), compared to the true kernel used to simulate the responses. **Bottom:** Correlation between the estimated and true kernel as a function of the number of trials. Shaded areas indicate variability across simulations (confidence intervals). As the number of trials increases, both methods improve, but the GLM consistently provides more accurate estimates, especially with fewer trials.

3.2.3.1 The double-pass procedure

The double-pass experiment is an experimental paradigm designed to isolate and quantify internal noise (A. J. Ahumada, 2002), distinguishing it from external noise or task-induced variability. By presenting identical stimuli twice on separate trials and analyzing response consistency, this method attributes any inconsistency to internal noise.

To avoid demand effects, the two presentations of identical stimuli are separated by many intervening trials (Hasan, Joosten, & Neri, 2012). For instance, in a two-alternative forced-choice (2AFC) task, the observer is presented with two stimuli (A and B) and their responses across repeated trials are categorized into four possibilities: AA, AB, BB, BA. These response types quantify decision consistency, with greater inconsistency (e.g., frequent AB or BA responses) indicating higher internal noise (Murray, Bennett, & Sekuler, 2002).

In practice, to estimate internal noise from double-pass data, we analyze two key behavioral metrics:

- Percentage of Agreement (p_{agree}): The proportion of trials where the participant gives the same response in both passes.

- Probability of Choosing Interval 1 (p_{first}): The proportion of trials where the participant selects response category 1 across all trials (in a 2IFC experiment; alternatively probability of choosing the first response option in 2AFC data).

From these empirical values, internal noise and decision bias are then estimated using grid search simulations. This method involves generating artificial observers that mimic human decision-making under different noise conditions. Each simulated observer follows the linear observer model, where responses are influenced by:

- A fixed decision criterion (bias, b) that determines the threshold for choosing one response over another.

- A level of internal noise σ_n that affects response variability.

In the simulation, a range of artificial observers is created, each with varying levels of internal noise and different response biases (favoring Interval 1 or 2). These simulated observers undergo a large number of trials, producing predicted values of (Pa, Pint1) for different noise and bias conditions. The next step involves comparing the participant's actual (Pa, Pint1) values against the simulated dataset. The algorithm identifies the simulated observer whose (Pa, Pint1) values best align with the participant's empirical data. This is achieved by minimizing the difference between the observed and simulated (Pa, Pint1) values. The internal noise (IN) and bias value that yield the closest match are selected as the final estimated parameters for the participant.

Figure 3.4 illustrates typical lookup data computed using a simulated linear observers with $\sigma_n \in [0,5]$ and $b \in [-5,5]$, expressed in units of external (stimulus) noise. Empirical probabilities p_{agree} and p_{first} are computed over simulated double-pass experiments with $n = 10^4$ repeated trials, and averaged over 10 realizations. For unbiased observers, σ_n maps non-linearly but bijectively to p_{agree}. However, symmetry of p_{agree} for positive and negative biases b illustrates the need to include p_{first} to disambiguate the underlying values of b and σ_n.

Palin Toolbox: Lookup simulation

Simulation of a lookup table using synthetic linear observers, systematically varying internal noise $\sigma_n \in [0,5]$ and decision criterion $b \in [-5,5]$ to compute the impact on agreement rate (p_{agree}) and bias (p_{first}). The resulting values are stored in a CSV file for later estimation from empirical data.

```
DoublePass.build_model(agreement_model_file='model_df.csv',
                       internal_noise_range=np.arange(0,5,.2),
                        criteria_range=np.arange(-5,5,1),
                     n_repeated_trials=1000, n_runs=10)
```

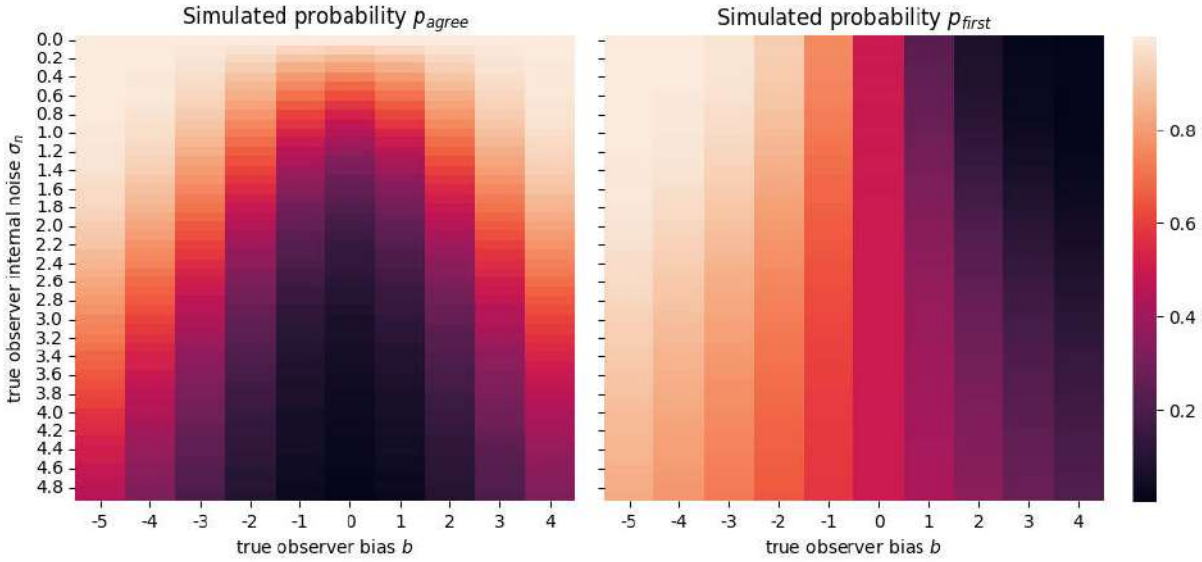


Fig. 3.4 **Lookup data computed using simulated linear observers** which show the influence of internal noise $\sigma_n \in [0,5]$ and decision criterion $b \in [-5,5]$ on p_{agree} and p_{first} S(forward model) **Left:** The simulated p_{agree} is highest when internal noise is low and bias is large (either positive or negative) and it is lowest when the bias is zero and internal noise is high. The plot is symmetric around $b = 0$, which illustrates the need of using p_{first} to disambiguate the direction of bias. **Right:** The simulated p_{first} increases monotonically with bias and captures directional trends in responses independently of stimuli. Lower internal noise and more positive bias are associated with a higher p_{first}.

Palin Toolbox: Internal Noise estimation

Creating 3 double-pass experiments (resp. with 150+150, 500+500, 1000+1000 trials), of the type 2AFC (Int2Trial), where stimuli are 6-dimensional, randomly sampled with a standard deviation of 100.

The simulation will have every observer meet every experiment, and on each response, run an InternalNoiseValue analyser. This Analyser takes 2 parameters: an internal noise estimation method (here, DoublePass).

When the simulation is run (`Simulation.run_all(n_runs)`), the Simulation will iterate over every configuration in `observer_params`, `experiment_params`, `analyser_params` and have every possible observer meet every possible experiment.

```
# Extract internal noise using the DoublePass method for one linear observer
DoublePass.extract_single_internal_noise(responses_df,agreement_model_file='agreement_model_large.csv')
# Simulate multiple linear observer to look at the Confidence interval of estimation
observer_params = {'kernel':['random'],'internal_noise_std':np.arange(0,5.1,0.5), 'criteria':[0]}
experiment_params = {'n_trials': [150,500,1000], 'trial_type': [Int2Trial],
    'n_features': [6],'external_noise_std': [100]}
analyser_params = {'internal_noise_extractor':[DoublePass],
                   'agreement_model_file':['agreement_model_large.csv']
sim_in = Sim(DoublePassExperiment, experiment_params, LinearObserver, observer_params,
          InternalNoiseValue, analyser_params)
sim_in_df = sim_in.run_all(n_runs=10)
```

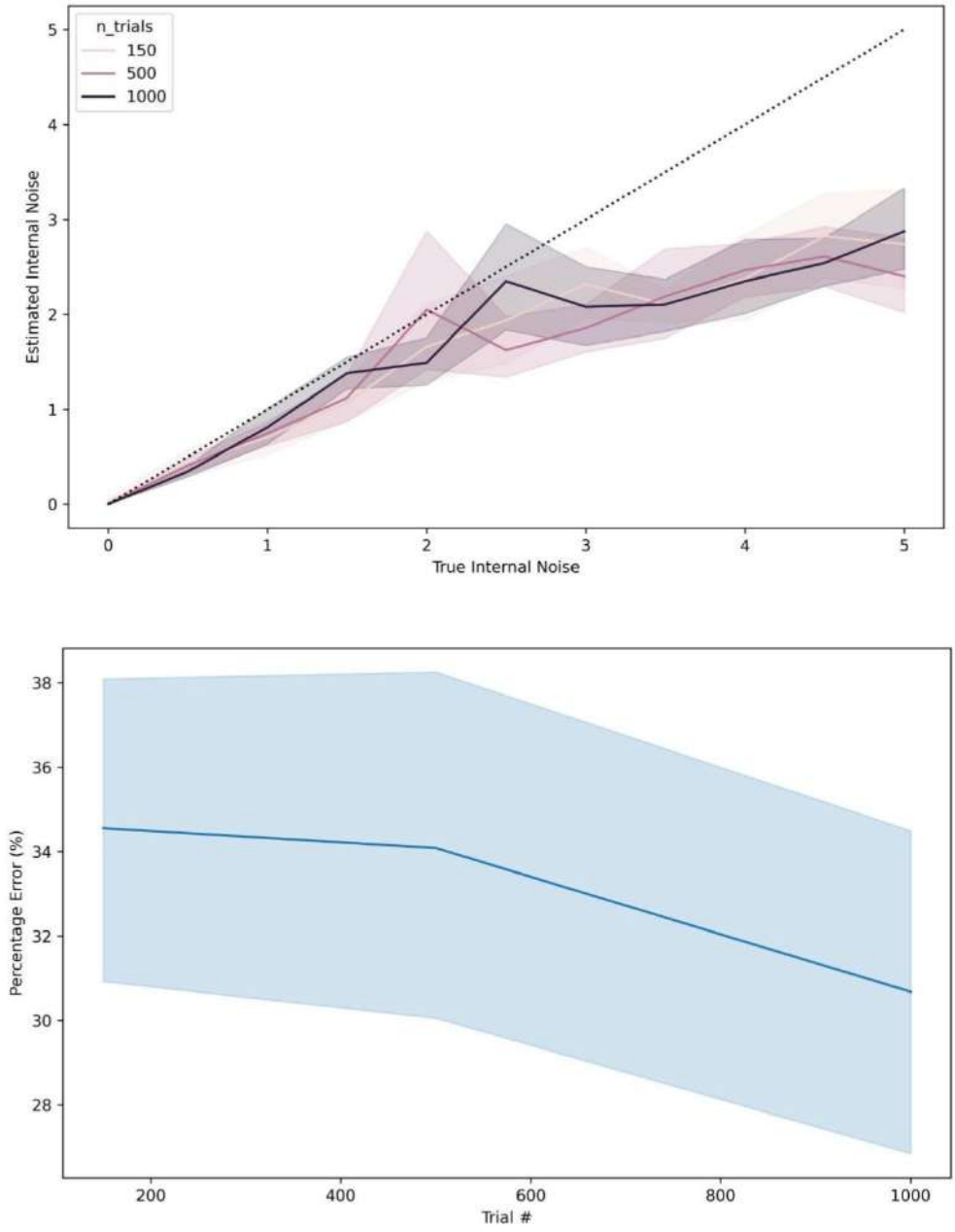


Fig. 3.5 **Effect of trial number on internal noise estimation accuracy. Top:** Estimated internal noise plotted against true internal noise values for three different trial counts (150, 500, and 1000). The diagonal line indicates perfect estimation. Accuracy improves with more trials, and estimates become more closely aligned with the ground truth up to approximately 3 standard deviations of internal noise. Beyond that point, the estimation begins to systematically underestimate higher internal noise values, regardless of the number of trials.**Bottom:** Percentage error in internal noise estimation as a function of the number of trials. The estimation error remains around 34% for trial counts between 200 and 600, but then decreases steadily, reaching around 20% with 1000 trials. This trend confirms that increasing the number of trials enhances estimation reliability and reduces variability.

3.3 Exploring prosody perception with reverse correlation methods

Reverse correlation methods, originally developed in sensory neuroscience, have recently been adapted to study high-level auditory representations such as speech prosody.

In particular, Ponsot, Arias, and Aucouturier (2018) introduced a reverse-correlation paradigm to access listeners' mental representations of interrogative intonation. In their experiment, five French speakers were presented with pairs of pitch-manipulated versions of the word *vraiment* ("really") and asked to decide which sounded more interrogative. The pitch contours were perturbed by applying a form of "intonation white noise", i.e., random pitch shifts sampled from Gaussian distributions, at six equally spaced time points (71 ms) across the utterance (Figure 3.6, Panel A), using a dedicated python toolbox (CLEESE, Burred, Ponsot, Goupil, Liuni, and Aucouturier (2018)). These manipulations preserved the identity of the word while randomizing its prosodic profile.

Participants' binary responses (1 if the utterance was judged as more interrogative, 0 otherwise) were then analysed with the classification-image method (Figure3.6, Panel B) to extract a kernel reflecting which parts of the pitch contour influence perceptual decisions.

The kernel (Figure 3.6, Panel C) showed a clear, consistent pitch rise toward the end of the word (segment 6), revealing that listeners systematically associate rising intonation in that region with interrogative intent. Importantly in Ponsot, Burred, et al. (2018), this pattern was replicated across all five healthy participants, who exhibited highly similar individual kernels, suggesting a shared internal representation of French interrogative prosody.

This approach shows how reverse correlation can uncover perceptual strategies without relying on explicit feature definitions, and also opens the opportunity to measure participant internal noise using the same procedure (something Ponsot, Burred, et al. (2018) didn't do for interrogative prosody). In the current thesis, we adopt a similar method to investigate prosodic processing (specifically, interrogative prosody) in stroke patients (see Chapter 4).

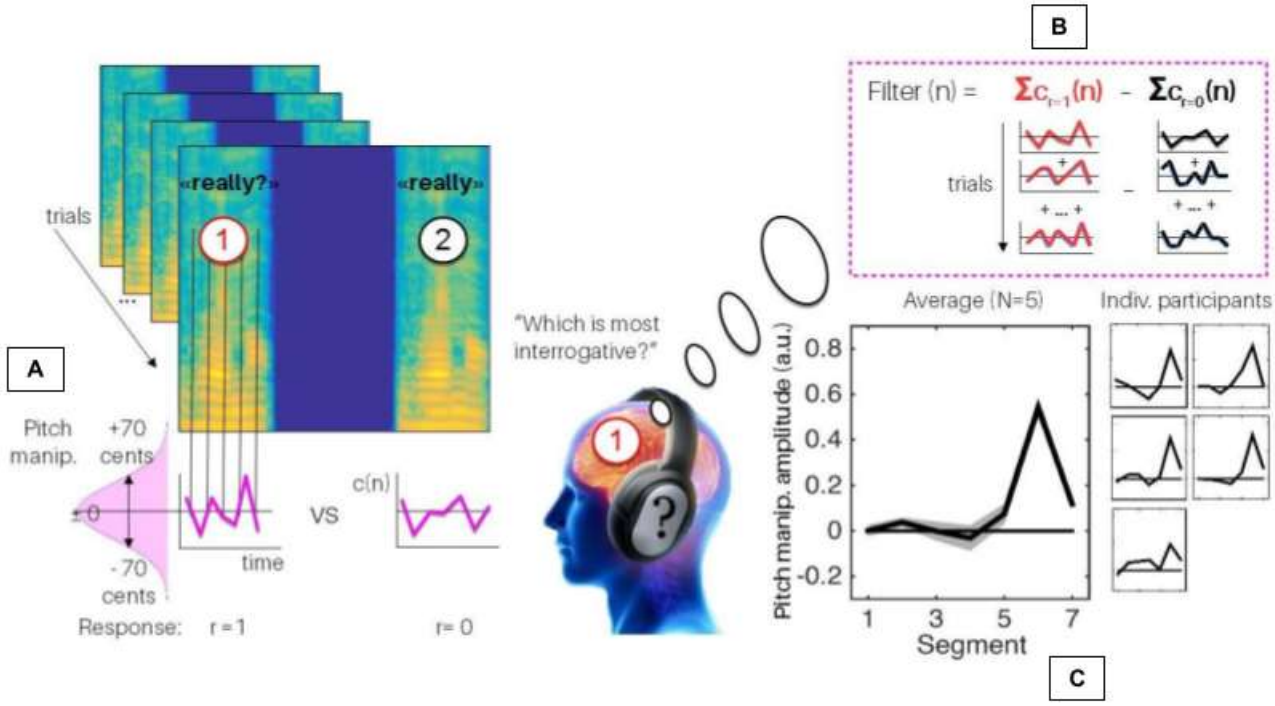


Fig. 3.6 **Reverse correlation paradigm for interrogative prosody (adapted from (Ponsot, Arias, & Aucouturier, 2018)) (A)** The pitch contour of the word *vraiment* ("really") is randomized across six segments using Gaussian-distributed perturbations, generating a variety of intonation patterns. On each trial, participants hear a pair of such manipulated utterances and choose the one that sounds more interrogative. **(B)** A prosodic kernel is computed by correlating responses with pitch contours across trials, effectively revealing how pitch modulations influence decisions. **(C)** The resulting kernel shows a consistent pitch rise at the end of the word, both at the group level (left) and across individual participants (right), indicating a shared mental representation of interrogative intonation in French.

3.4 Discussion

In Chapter2, we explored the biological basis of prosody perception, emphasizing its critical role in communication and social interaction after stroke. We also discussed the challenges posed by traditional assessment tools, which often fail to provide mechanistic and specific insights into deficits like aprosodia. These limitations highlight the need for approaches that move beyond descriptive measures to quantify prosody perception in a systematic and theory-driven way.

In Chapter3, we took a step toward addressing these gaps by introducing possible mechanistic model rooted in computational psychophysics. Specifically, we examined

psychophysical experiments such as reverse correlation to probe how participants process and represent prosodic cues. This framework allowed us to compute key variables:

- Internal representations, estimated using classification images, to map the auditory features participants rely on for prosody judgments.
- Internal noise, quantified through double-pass experiments, to measure variability in the perceptual and decision-making processes.

These methods provide insights into what model may interpret prosody perception better by offering precise, computationally grounded metrics for understanding how individuals perceive and respond to prosodic stimuli.

Next chapter4 lays the groundwork for transitioning from theoretical exploration to empirical application. By combining the mechanistic models from Chapter 2 with real-world data, we aim to develop a framework that bridges theory, data, and computation. Computational psychophysics provides a powerful lens through which to study prosody perception, enabling us to:

- Apply models to actual data, testing their predictive power and robustness in quantifying deficits.
- Explore individual differences in prosody perception, paving the way for personalized assessment tools.

Part II

A first encounter with clinical data: preliminary results and problem statement

CHAPTER 4

Reverse-correlation biomarkers of aprosodia after a right-hemisphere stroke

As discussed in Chapter 2, survivors of right-hemisphere stroke may experience deficits in prosody perception that persist from the acute to chronic phase. These impairments can often remain subtle, apparent only in conversations with family members while going undetected during speech therapy sessions when assessed using standard batteries. Reverse correlation, the approach sketched out in Chapter 3, appears to be a promising method to uncover the psychophysical parameters of such deficits, when present.

In part II, we propose a first *encounter* with clinical data. In the present chapter, we present an analysis of reverse correlation data, collected prior to the thesis (the work of speech therapy students Mélissa Jeulin, Pauline Bardet, Pauline Commère, supervised by Marie Villain and JJ Aucouturier), using the reverse correlation paradigm of Ponsot, Burred, et al. (2018). Our goal is to implement reverse correlation analysis in this context, extract linear-observer parameters and compare their diagnostic value to existing aprosodia scales commonly used in speech therapy. As we'll see, this study will also give us the opportunity to identify computational limitations with the classical kernel and internal

noise extraction method that occur when analysing stroke patients (Chapter 5), which the rest of the thesis (Part III) will set to address. Part IV will provide a re-analysis of the same data, using our new proposed methods.

This study was published as:

> Adl Zarrabi, A., Jeulin, M., Bardet, P., Commère, P., Naccache, L., Aucouturier, J. J. & Villain, M. (2024). *A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after right-hemisphere stroke*. Scientific Reports, 14(1), 15194.

We present here a verbatim of the manuscript, preceded by a short summary of the methods and main results.

4.1 Materials and methods

We adapted the paradigm of Ponsot, Burred, et al. (2018) to probe the sensory/cognitive mechanisms that underlie the processing of interrogative prosody in stroke patients. To do so, we recorded the utterance of the French word "vraiment" ("really") and generated prosodic variations by introducing random Gaussian noise to the base sound for each stimulus. The utterance was segmented into six intervals of 71 ms, and the pitch at each breakpoint was independently manipulated using a normal distribution, using the CLEESE python toolbox (Burred et al., 2018). Participants were then asked to judge between two random prosodic variations in each 2IFC trial, identifying which interval contained the word that sounded most interrogative.

To extract participants' linear-observer parameters (kernels k and internal noise σ_n), we used the classical methods from the reverse-correlation literature. Kernels were extracted using the "classification-image"/weighted-sum method of averaging the noisy stimulus from stimuli where participants provided negative responses and subtracting it from the average noisy stimulus of stimuli with positive responses (Murray, 2011). This process results in a weighted sum that reflects participants' internal representation of the interrogative form of "vraiment".

Internal noise was estimated using the double-pass procedure (Burgess & Colborne,

1988): the experiment consisted of 150 trials, separated into 3 blocks of 50 trials, and the second and third blocks were exactly repeated trials. We estimated observer consistency by calculating the percentage of agreement in their responses across identical trials and estimated the standard deviation of the equivalent internal noise using a lookup table, as detailed in Chapter 3.

Using this procedure, we compared data from right-hemisphere stroke patients (N=22) and age-matched healthy controls (N=21) to establish a baseline for the normal prosody perception of the interrogative form of " vraiment?", which we hypothesised from Ponsot, Burred, et al. (2018) and others to include a pitch rise at the end of the utterance (Banuazizi & Creswell, 1999). This comparison allows us to define a reference threshold for typical performance and identify abnormalities in prosody processing. Additionally, we collected patient responses to various standardized assessment batteries used in speech therapy sessions, serving as pathological gold standards. These included the MEC Comprehension and Repetition tests, Airtac2, LAMA, and MBEA, HADS (see below), which provided clinical benchmarks for evaluating the "biomarker potential" of kernel and internal noise.

4.2 Results

Both kernel and internal noise extracted from the reverse-correlation procedure effectively distinguished patients from controls.

Internal representations of interrogative prosody in the control group exhibited a typical final-rise contour, characterized by a marked pitch increase at the end of the second syllable. In contrast, patients' internal representations had lower amplitude, indicating reduced discriminative power, and displayed greater variability across individuals .

Additionally, control participants demonstrated high response consistency across trials, with internal noise values ($M = 0.7$, $SD = 0.37$) within the range typically observed for lower-level auditory and visual tasks (Neri, 2010). Patient responses were associated with significantly higher levels of internal noise ($M = 2.54$, $SD = 1.90$).

Statistical analyses confirmed significant differences between groups for both representation typicality - the similarity between an individual's mental representation and the

control group average - $M = 0.27$ [0.16; 0.39], Mann–Whitney's $U(-0.82) = 420$, $p < 0.001$ and internal noise $M = -1.84$ [−2.61; −1.07], $U(0.59) = 95.00$, $p = 0.001$.

Within the patient group, internal noise values and, to a lesser extent, representation typicality were statistically associated with scores from the current gold standard for assessing prosody perception deficits (MEC), demonstrating good concurrent validity. Higher internal noise values correlated with lower (more severe) scores on the MEC *prosody comprehension* scale ($p = 0.043$), while representation typicality showed a positive but non-significant trend ($p = 0.15$). Notably, neither measure correlated with the MEC *prosody repetition* score ($p = 0.82$ and$p = 0.365$, respectively), despite a significant positive correlation between the two MEC scores. This highlights the symptom specificity of our measures.

A well-known limitation of the MEC instrument is its poor sensitivity, as some patients scoring above the pathological cut-off (9/12) still report communication difficulties. Our measures, however, successfully distinguished this group of MEC-negative (high-functioning) patients from controls, both in terms of representation typicality ($p = 0.001$) and internal noise ($p = 0.026$).

To further examine the convergent validity and specificity of internal representation and internal noise measures, we explored their associations with other constructs relevant to stroke rehabilitation. As expected, both measures correlated with non-prosody-related difficulties in tone intensity and duration discrimination, as assessed by AIRTAC2 ($p = 0.007$ and $p = 0.037$, respectively). However, they were not associated with the ability to detect rare auditory targets among distractors (LAMA, $p = 0.25$ and $p = 0.23$) or with musical melody processing, as measured by MBEA ($p = 0.46$ and $p = 0.98$).

Regarding musical ability, MBEA was assessed in a subset of patients, the majority of whom were found to have deficits in melody/pitch processing. Among patients classified with melody amusia, 75% exhibited internal representations that visually deviated from controls, whereas 60% of non-amusic patients had normal representations. Finally, internal noise, but not representation typicality, was significantly related to anxiety and depression levels, as measured by HADS ($p = 0.018$ and $p = 0.178$, respectively).

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after right-hemisphere stroke

Aynaz Adl Zarrabi[1,5], Mélissa Jeulin[2,5], Pauline Bardet[2], Pauline Commère[2], Lionel Naccache[2,3], Jean-Julien Aucouturier[1], Emmanuel Ponsot[4] & Marie Villain[2,3]✉

After a right hemisphere stroke, more than half of the patients are impaired in their capacity to produce or comprehend speech prosody. Yet, and despite its social-cognitive consequences for patients, aprosodia following stroke has received scant attention. In this report, we introduce a novel, simple psychophysical procedure which, by combining systematic digital manipulations of speech stimuli and reverse-correlation analysis, allows estimating the internal sensory representations that subtend how individual patients perceive speech prosody, and the level of internal noise that govern behavioral variability in how patients apply these representations. Tested on a sample of N = 22 right-hemisphere stroke survivors and N = 21 age-matched controls, the representation + noise model provides a promising alternative to the clinical gold standard for evaluating aprosodia (MEC): both parameters strongly associate with receptive, and not expressive, aprosodia measured by MEC within the patient group; they have better sensitivity than MEC for separating high-functioning patients from controls; and have good specificity with respect to non-prosody-related impairments of auditory attention and processing. Taken together, individual differences in either internal representation, internal noise, or both, paint a potent portrait of the variety of sensory/cognitive mechanisms that can explain impairments of prosody processing after stroke.

Keywords Stroke, Prosody, Reverse-correlation, Internal noise, Perception

After a right hemisphere stroke, more than half of the patients present a communication disorder such as aprosodia, the impossibility to produce or comprehend speech prosody—or the "melody" of speech[1–5]. Despite the social-cognitive implications for patients of not being able to process e.g. linguistic or emotional prosody[6], aprosodia following stroke has received scant attention.

First, the existing assessment tools for impairments of prosodic processing are found to be lacking in several aspects. The gold standard in the French language, the "Montréal Evaluation de la Communication" (MEC)[7] consists of a combination of listening and production tests which exhibit good inter-rater reliability but are suspected of limited sensitivity, failing to capture nuanced deficits in language processing in e.g. ecological situations[8]. More generally, traditional pre-post assessments with listening batteries (ex. the 12-items of the MEC prosody task) suffer from test–retest effects, where participants might remember their responses, leading to learning effects. Additionally, assessments based on prosody production typically involve manual scoring by clinicians, which may generate issues of inter-rater variability and limits the potential for monitoring patients remotely. Finally,

[1]Université de Franche-Comté, SUPMICROTECH, CNRS, Institut FEMTO-ST, 25000 Besançon, France. [2]Department of Physical Medicine & Rehabilitation, APHP/Hôpital Pitié-Salpêtrière, 75013 Paris, France. [3]Paris Brain Institute (ICM), Inserm, CNRS, PICNIC-Lab, 75013 Paris, France. [4]Science & Technology of Music and Sound, IRCAM/CNRS/Sorbonne Université, 75004 Paris, France. [5]These authors contributed equally: Aynaz Adl Zarrabi and Mélissa Jeulin. ✉email: marie.villain@aphp.fr

existing tools typically provide a binary score indicating the presence or absence of a pathology, but do not allow for an in-depth understanding of the mechanisms that explain why a specific error may occur.

Besides lacking sensitive assessment tools, the field is also lacking in its understanding the exact sensory/cognitive mechanisms that subtend aprosodia[4]. On the one hand, a wealth of cognitive neuroscience research has linked linguistic and/or emotional prosody perception with a dominantly-right temporo-frontal network[9]—although it should be noted that recent research has also implicated a wider variety of cortical and subcortical networks[10]. One prominent explanation for such a specialization proposes that the bilateral auditory cortices differ in their temporal and spectral resolution, with left auditory regions responding preferably to fast changes in the type of spectral cues implicated in phonetic discrimination, and right auditory regions to slow variations of pitch as seen in speech prosody and music[11]. On the other hand, clinical patient data has also linked right hemisphere damage due to stroke with a wide multitude of cognitive-communication deficits, which not only include aprosodia, but also impairments of the interpersonal communication such as inappropriate pragmatics and humour[1], as well as domain-general deficits in attention, memory and executive function[5]. It therefore remains poorly understood whether impairments of prosody perception result from specific damage in regions involved in speech representations, or in more generic mechanisms[4]. Lacking a mechanistic understanding of why patients perform poorly on such tasks deprives health practitioners of practical therapeutic targets for their subsequent rehabilitation.

When studying the neural mechanisms that relate physical stimuli to perception, the modern field of psychophysics has largely moved from simply measuring sensory thresholds and psychometric functions, and now provides a toolbox of techniques to measure and fit multi-staged models able to simulate participant behaviour[12]. Notably for the example of speech prosody, the psychophysical technique of reverse-correlation (or "classification images")[13] allows estimating, at the individual level, not only what sensory representations subtend the normal or abnormal perception of e.g. interrogative prosody[14], but also "internal noise" parameters that capture aspects of behavioral variability that are of potential neurological relevance[15,16].

While the representation + noise model has a rich history in healthy participants, with or without peripheral hearing impairment[17,18], its use in participants with neurological or developmental disorders has received relatively little attention[19–21]. Here we show on a sample of N = 22 right-hemisphere brain stroke survivors that such simple procedures promise to enrich the current clinical toolbox with more sensitive and informative markers of receptive aprosodia. While the same tool can be applied to study a variety of prosodic functions, incl. social or emotional, in this study we specifically target the perception of linguistic prosody, defined as the acoustic variations of suprasegmental cues such as tone, amplitude and speech rate that support language analysis beyond the phonetic level, incl. syntax, semantics, and discourse structure[22]—for instance shaping sentences into questions or statements with rising or falling intonations[23]. Using reverse correlation, we show that it is possible to estimate not only the internal representations that subtend how individual patients specifically perceive interrogative prosody, but also a quantitative measure of the consistency with which patients apply these representations in perceptual tasks, and that these two parameters have potential to surpass both the sensitivity and diagnostic richness of existing tools.

Materials and methods

Participants

N = 22 brain stroke survivors (male: 17; M = 57 yo, SD = 12.43), and N = 21 age-matched controls (male: 13; M = 58 yo, SD = 13.34) took part in the study. There was no significant sex distribution difference between groups (Chi-square test, p = 0.368), and no significant age difference (Mann–Whitney, p = 0.970).

All patients were in- or out-patients of the Physical Medicine & Rehabilitation Department, APHP Pitié-Salpêtrière Hospital in Paris, France, undergoing speech therapy for different deficits post-stroke like swallowing difficulties, neuro-visual impairments, attentional impairments, neglect, dysphasia etc. Patients included in the study (Table 1) had a history of supratentorial right-hemisphere ischemic stroke, corroborated by clinical assessments NIH stroke scale (NIHSS; M = 10.8) and brain MRI, and dating less than 1y (Median = 4 months) at the time of inclusion; were first-language French speakers; and had no disorders of wakefulness/consciousness, dementia, severe dysarthria, psychiatric antecedents (> 2 months in-patient) or major visual or auditory impairment (> 40 dB HL). Patients with language comprehension deficits -aphasia- (score < 10/15 on the BDAE instruction-following task) were excluded from the study.

In addition, we recruited a group of N = 21 controls matched in age, sex and degree of hearing loss. Seven of these control participants were recruited via the INSEAD-Sorbonne Université Center for Behavioral Science, Paris, France, and took part in the experiment in a laboratory setting. The remaining 15 were recruited among the FEMTO-ST participant pool, and took part in an online version of the same procedure. Among these 15 online participants, we concluded that one participant was not sufficiently engaged in the task, statistical results conducted with the full control group (including this outlier) are qualitatively similar to our main text conclusions, and presented in Supplementary Text 2.

Clinical assessment

Two subtests of the French version of the "Montréal Evaluation de la Communication" (MEC) protocol[7] were administered to the patients to assess their linguistic prosody capacities (comprehension and repetition). The linguistic prosody comprehension subtest evaluated the ability to identify linguistic intonation patterns. This subtest consists of four semantically neutral simple sentences and each one is presented to the patient with three different intonations, for a total of 12 items. After listening to a sentence, the patient is asked to select the correct intonation among the three different written options (interrogative, imperative or affirmative). The linguistic prosody repetition subtest examines the ability to verbally reproduce linguistic intonations. It is formed of the

		Controls	Patients
n		21	22
Sex, n (%)	f	8 (38.1%)	5 (22.7%)
	m	13 (61.9%)	17 (77.3%)
Age, median [min, Q1, Q3, max]		58 yo [27,52,64,82]	60.5 yo [28,52.2,63,74]
Month after stroke, median [min, Q1, Q3, max]			4 mo [0,1,5,17]
Stroke type, n (%)	HEM		3 (33.3%)
	ISCH		6 (66.7%)
NIH stroke scale (NIHSS), median [min, Q1, Q3, max]			Available: N = 11(50%) 10 [2,5.5,16,20]
MEC Prosody Comprehension item, median [min, Q1,Q3, max]			Available: N = 22 (100%) 9 [0,8,11,12]
MEC Prosody Repetition item, median [min, Q1, Q3, max]			Available: N = 22 (100%) 11 [7,10,12,12]
MEC Total, median [min, Q1, Q3, max]			Available: N = 22 (100%) 21 [9,18.2,22.8,24]
BDAE command execution item, median [min, Q1, Q3, max]			Available: N = 22(100%) 14 [5,14,15,19]
Audiogram left-ear, median dBHL at 1000 Hz [min, Q1, Q3, max]		0 dBHL [0,0,15,35]	Available: N = 7(31%) 20 dBHL [10,15,30,60]
Audiogram right-ear, median dBHL at 1000 Hz [min, Q1, Q3, max]		5 dBHL [0,0,20,30]	Available: N = 7(31%) 15 dBHL [5,7.5,37.5,45]
Vocal audiogram, median% detection at 40 dB [min, Q1, Q3, max]			Available: N = 13(59%) 99. % [85,94,100,100]
LAMA Sustained auditory attention score accuracy, median [min, Q1, Q3, max]			Available: N = 12(54%) 30 [29,29.8,30,30]
LAMA Sustained auditory attention reaction time (sec), median [min, Q1, Q3, max]			Available: N = 12(54%) 92.5 [63,85.8,137,192]
MBEA (Montreal Battery of Evaluation of Amusia), median [min, Q1, Q3, max]			Available: N = 13(59%) 60 [48,57,71,85]
AIRTAC2 (Auditory discrimination), median [min, Q1, Q3, max]			Available: N = 13(59%) 44 [36,42,47,48]
HADS (depression + anxiety), median [min, Q1, Q3, max]			Available: N = 13(59%) 18.5 [7,11.2,24.8,35]

Table 1. Patients and control demographics and clinical characteristics. N = 22 right-hemisphere stroke survivors and N = 22 age-matched controls took part in the study. MEC, Montréal Evaluation de la Communication; BDAE, Boston Diagnostic Aphasia Examination.

same four sentences as the comprehension task. The previously recorded stimuli are presented in random order. The patient is asked to repeat each sentence with the same intonation. The maximum score is 12 for both subtests.

In order to exclude patients with a significant hearing impairment from the study, patients were assessed using Lafon's cochlear lists of monosyllabic words (List 2 and List 3)[24]. These were calibrated at an intensity of 40 decibels (dB) and played through headphones. Only patients who scored 80% or more on both lists were included. In addition, the Boston Diagnostic Aphasia Examination (BDAE) command execution subtest[25] was used to exclude patients with comprehension disorders. Only patients with a score of 12/15 or higher were included. Some patients underwent MMSE (Mini-Mental State Examination) or MoCA (Montreal Cognitive Assessment) evaluations as part of their clinical follow-up, but these assessments were conducted at different times, making it difficult to perform direct comparisons. It's important to note that the Boston Diagnostic Aphasia Examination (BDAE) command execution subtest[25] was used to exclude patients with comprehension disorders. Only patients with a score of 12/15 or higher were included. And none of these patients suffered from aphasia, as it was an exclusion criterion for our study.

To assess possible mood disorders, the Hospital Anxiety and Depression Scale (HADS) self-questionnaire[26] was administered to patients to assess their current level of anxiety and depression. It contains 7 questions for the anxiety part and 7 questions for the depression part, with a separate score for each. A score of 11 or more for each part indicates a possible anxiety and/or depression state.

To assess auditory attention, a subset of patients also underwent the sustained auditory attention subtest of the "Logiciel d'Attention en Modalité Auditive" (LAMA)[27]. The assessment and rehabilitation software "Aide Informatisée pour la Rééducation des Troubles Auditifs Centraux " (Airtac2)[28] was used to assess central auditory processing. Intensity discrimination and duration discrimination of non-verbal sounds were proposed to compare central auditory processing abilities with the results of the Reverse Correlation task. Finally, the Montreal Battery of Evaluation of Amusia (MBEA)[29] was selected to assess the music perception abilities of a subset of patients. Since the disorder of music perception (amusia) is primarily a disorder of pitch perception[30], the three tasks in the melodic organization part (scale test, contour test, interval test) were selected (See Table 1 for details).

Procedure

We recorded a 426-ms utterance of the French word "vraiment" ("really"), and generated prosodic variations by dividing it into six segments of 71 ms and randomly manipulating the pitch of each breakpoint independently using a normal distribution (SD = 70 cents; clipped at ± 2.2 SD), hereafter referred to as "stimulus noise". These values were linearly interpolated between time points and fed to an open-source pitch-shifting toolbox (CLEESE, Python language, v1.0, available at https://github.com/neuro-team-femto/cleese) developed for this purpose[31]. We then presented patients with 150 successive pairs of such manipulated utterances (*really/really?*) asking them to judge which, within each pair, sounded most interrogative (examples of sound stimuli are available in the code repository shared with the article—see *Code Availability*). The sequence was divided into 3 blocks of 50 pairs. Without the participant's knowing, the first and last block of each sequence contained identical pairs of sounds (a procedure called double-pas[15,32], allowing us to examine response variability), but all other sounds in the sequence were otherwise distinct (in more details, N = 9/22 patients and N = 7/21 controls had only 25 repeated trials among block 2 and 3, while the other N = 13/22 patients and N = 14/21 controls had a complete repetition of the 50 trials in blocks 2 and 3; there was no statistical difference between the levels of internal noise measured with these two setups (patients: Mann–Whitney p = 0.24; controls: p = 0.13). N = 9/22 patients were additionally tested 4 repeated (one week apart), but we have only retained the first session and did not include these extra data points in the statistical analysis. Sounds were delivered using closed headphones (Beyerdynamics DT770) presented the stimuli dichotically (same signal in both ears) at an identical comfortable sound level (~ 70 dB SPL) to all patients and healthy subjects. The inter-stimulus interval in each pair was 500 ms, and the interval between successive pairs was 1 s. The procedure took about 15 min to complete.

Reverse-correlation analysis

For each participant's response data, we fitted a 2-stage psychophysical model consisting, first, of a prosodic template (or "internal representation") to which sound stimuli are compared and, second, of a level of "internal noise" which controls how consistently this representation is applied to incoming stimuli (Fig. 1).

Participants' internal representations (a time × pitch representation of an ideally interrogative pitch contour) were computed using the classification image technique[13] to differentiate between interrogative and non-interrogative pitch contours. Specifically, we subtracted the average pitch contour of non-interrogative classifications from that of interrogative classifications. To normalize this resultant representation, we divide it by the root mean square of its values—this method involves squaring each value of the representation, averaging these squared numbers, and then taking the square root of this average to scale the representation accordingly. For each patient,

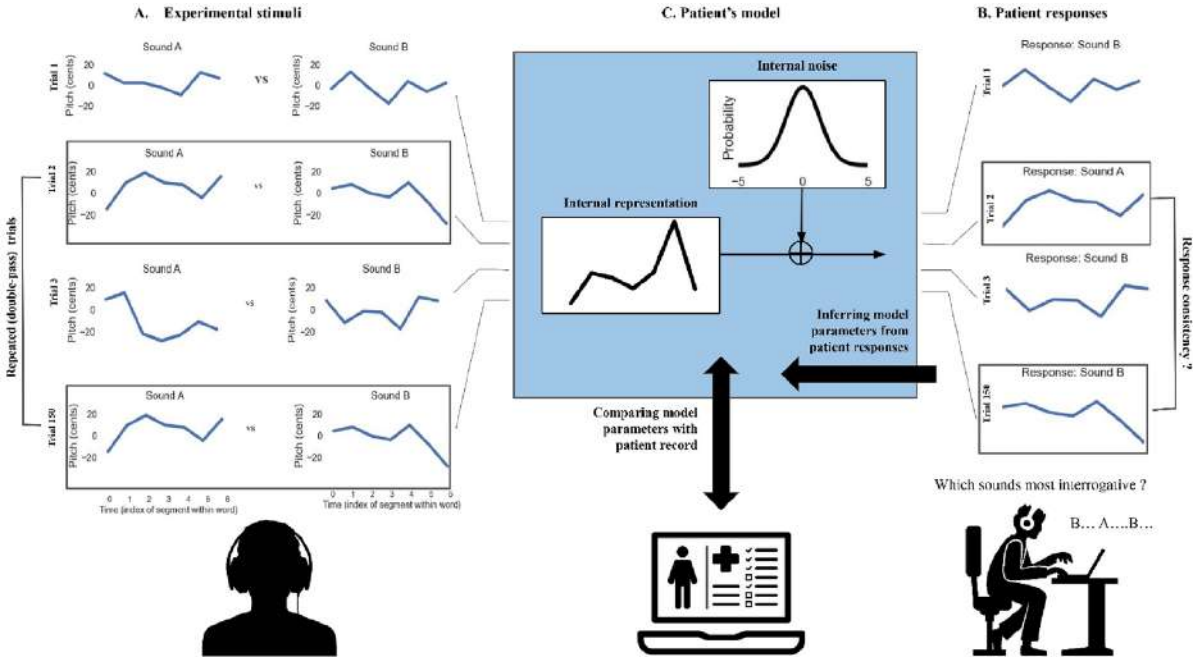


Figure 1. The representation + noise model. Patients were presented with 150 successive trials consisting of pairs of manipulated prosodies (**A**) and asked to judge, within each pair, which sounded most interrogative (**B**). Patient responses in each trial were fitted with a 2-stage psychophysical model (**C**), consisting, first, of a prosodic template (or "internal representation") to which sound stimuli are compared and, second, of a level of "internal noise" which controls how consistently this representation is applied to incoming stimuli. See main text for details about the model-fitting procedure. In this work, we estimate the two model parameters (representation and noise) for each patient individually and compare them with patient records to test their value as markers of receptive aprosodia.

we then quantified how similar their internal representation is to the average representation in the control group, by computing the mean squared error between the two representations, and used this "representation typicality" as a parameter to correlate with clinical measures. Representations for controls were computed using the same procedure, using only the first 150 trials of each session in order to match the number of trials seen by patients.

Participants' internal noise (expressed in units of the standard deviation of stimulus noise) was inferred from response consistency and response bias across the repeated double-pass trials, using the simulation procedure of Neri[15]. In short, we computed an idealized participant model responding to repeated stimuli pairs of various sensory evidence, perturbed its response with additive gaussian noise ("internal noise"), and estimated the probability for that model to give the same response for identical trials (i.e. response consistency) and the probability of giving the first response option (i.e. response bias), for different standard deviations of that internal noise. For each participant, we then inverted that model and obtained the value of internal noise (by exhaustive search between 0 and + 5 std) that minimized the error between the observed and predicted values for that participant's consistency and bias. As in previous studies[15], we estimated internal noise conservatively between [0; + 5 std] in order to avoid unreliable estimates at large values, a known problem with double-pass procedures (see Appendix A). Internal noise values in the upper side of that range (e.g. illustrated in Fig. 3 between 4.8 and 5) may either correspond to true internal noise values, or to larger values for which we could not provide an exact estimate.

Both of these analyses (internal representations and internal noise) were conducted using an open-source Python toolbox built for this purpose (PALIN v1.0, Python language, v1.5, available at https://github.com/neuro-team-femto/palin).

Statistical analysis

Group comparisons: because distributions of representation typicality and internal noise scores between patients and controls were non-normal, we compared population means using non-parametric (Mann–Whitney) independent sample t-tests.

Correlation with clinical measures: linear associations between representation typicality and internal noise, and clinical assessments (MEC, Prosody Comprehension, Prosody Repetition, Airtac2) met the homoskedasticity assumption and were therefore estimated using ordinary least-square regressions without robust (HC) norms, as these are considered to increase false positive rates when testing small samples. In addition, because regression residuals were occasionally non-normal, we estimated statistical significance using bootstrapped confidence intervals[33]. The analysis was implemented with the pymer.lm package[34] v4 0.8.2.

Ethics statement

The study was approved by *Comité de Protection des Personnes* CPP Ile-De-France V (ProsAVC, Decision of 22/07/2020). All methods in this study were carried out in accordance with the relevant guidelines and regulations, and all data in this study were obtained with informed consent from all subjects and/or their legal guardian(s).

Results

Both measures extracted from the reverse-correlation procedure allowed separating patients from controls: internal representations of interrogative prosody computed from control group responses exhibited a typical final-rise contour[14], with a marked increase of pitch at the end of the second syllable (Fig. 2-left), and control participants were able to apply these representations remarkably consistently across trials, with internal noise values M = 0.7 (SD = 0.37) in the range of those typically observed for lower-level auditory and visual tasks[15] (Fig. 2-right). In contrast, patients' internal representations had both lower amplitude (indicating less discriminative power) and more variable shape across individuals (Fig. 2-left; see also Fig. 3), and were applied with higher levels of internal noise (M = 2.54, SD = 1.90; Fig. 2-right). The two groups differed statistically for both representation typicality: M = 0.27 [0.16; 0.39], Mann–Whitney's U(− 0.82) = 420, $p < 0.001$; and internal noise: M = − 1.84 [− 2.61; − 1.07], U(0.59) = 95.00, p = 0.001.

Within the patient group, internal noise values (and, to a lower extent, representation typicality) were statistically associated with scores of the current gold standard for assessing deficits of prosody perception (MEC), demonstrating good concurrent validity. First, larger internal noise values were associated with lower (more severe) scores on the MEC prosody comprehension scale: noise: $R^2 = 0.189$, β = − 0.303 [− 0.596; − 0.010], t(20) = − 2.158, p = 0.043. Representation typicality also improved with better scores, albeit non-statistically ($R^2 = 0.100$, β = + 0.03 [− 0.012; + 0.071], t(20) = 1.49, p = 0.15). Second, both measures had also good symptom specificity, as strikingly neither correlated with the MEC score for prosody repetition (representation: $R^2 = 0.002$, t(20) = − 0.219, p = 0.82, noise: $R^2 = 0.041$, t(20) = − 0.92, p = 0.365), while both MEC scores were themselves positively correlated (r = 0.53).

An oft-quoted limitation of the MEC instrument is its poor sensitivity, with patients above the pathological cut-off on the MEC prosody comprehension scale (9/12) still complaining of communication difficulties[8]. Interestingly, our measures allowed clear separation of this group of MEC-negative patients (i.e. patients with MEC > 9) (N = 12/22) and controls (N = 21), both in terms of typicality of representation (M = 0.18 [0.06; 0.32], U (− 0.74) = 219.0, p = 0.001) and internal noise (M = − 1.54 [− 2.62; − 0.53], U(0.48) = 66, p = 0.026).

Finally, to examine the convergent validity and specificity of internal representation and internal noise measures, we investigated whether they were statistically associated with other constructs linked to central deficits common in stroke rehabilitation. Expectedly, both measures were associated with difficulties discriminating tone intensity and tone duration, as measured by AIRTAC2 (representation: R^2: 0.49, β = + 0.040 [0.013; 0.068], t(11) = 3.27, p = 0.007; noise: R^2: 0.33, β = − 0.28 [− 0.54; − 0.020] t(11) = − 2.36, p = 0.037). However, they were not associated with the patient's capacity to detect rare auditory targets among distractors, as measured by LAMA

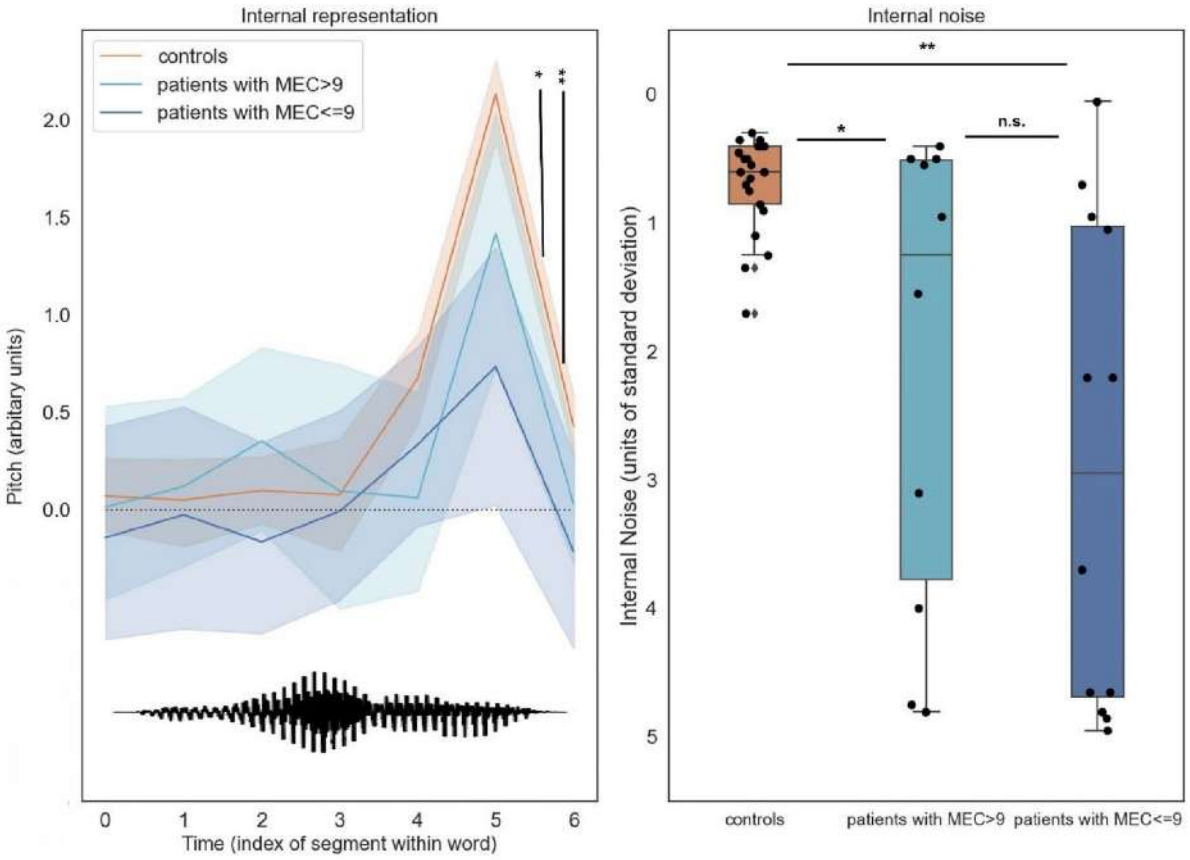


Figure 2. Patient parameters (internal representations and internal noise) estimated by reverse-correlation separate controls from patients above and below the pathological cut-off on the MEC prosody comprehension scale (9/12). Left: Internal representations of interrogative prosody computed from control group responses exhibited a typical final-rise contour, with a marked increase of pitch at the end of the second syllable. In contrast, patients' internal representations had both lower amplitude and more variable shape across individuals. The bottom waveform illustrates the shape of the base sound used to generate stimuli (a male-recording of the word vraiment/really). Right: control participants were able to apply these representations remarkably consistently across trials, with internal noise values < 1 standard deviations of stimulus noise. In contrast, patients' internal noise levels were larger and more variable, and scaled with prosodic difficulties measured by MEC.

(representation: R^2: 0.130, t(10) = 1.22, p = 0.25; noise: R^2: 0.136, t(10) = − 0.125, p = 0.23); or with the patient's capacity to process musical melodies, as measured by MBEA scale and melody items (representation: R^2: 0.050, t(11) = 0.765, p = 0.46; noise: R^2: 0.00, t(11) = 0.027, p = 0.98). Regarding music ability in particular, MBEA was assessed in N = 13 (59%) of our patients, the majority of which N = 8 (62%) were found impaired for melody/pitch processing with scores below the pathological cut-off score of 65/90. Out of the 8 patients who scored with melody amusia, 6 (75%) had representations that visually departed from controls. Comparatively, 3 out 5 (60%) of the patients without amusia had normal representations (Fig. S2).

Finally, internal noise (but not representation typicality) was found related to patients' level of anxiety and depression, as measured by HADS (noise R^2 = 0.249; β = 0.108 [0.021; 0.196], t(20) = 2.57, p = 0.018; representation R^2 = 0.089, t(20) = − 1.39, p = 0.178).

Discussion

In this report, we introduced a novel, simple psychophysical procedure which, by combining systematic digital manipulations of speech stimuli and reverse-correlation analysis, allows estimating the internal sensory representations that subtend how individual patients perceive speech prosody, as well as the level of internal noise that govern behavioral variability in how patients apply these representations in prosodic perceptual tasks.

Tested on a sample of N = 22 right-hemisphere stroke survivors, our two proposed parameters of representation typicality and internal noise provide a promising alternative to the clinical gold standard for evaluating impairments of prosody processing (MEC). First, internal noise (and, to a lesser extent, internal representations) strongly associate with receptive aprosodia, and not expressive aprosodia, measured respectively by MEC recognition and repetition scores within the patient group. Second, internal representations (and, to a lesser extent,

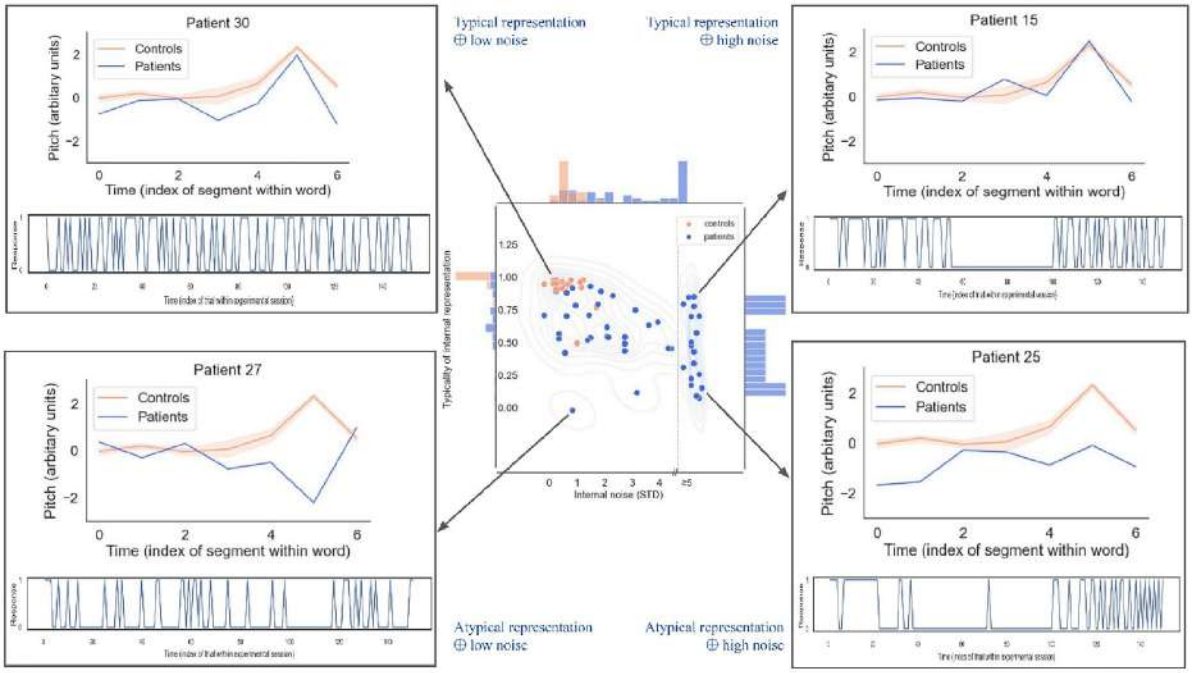


Figure 3. The representation + noise model captures a rich diversity of sensory/cognitive mechanisms underlying impairments of prosody processing after stroke. Center: Distribution of representation typicality and internal noise for controls and patients (considering all 4 sessions), overlaid with by kernel density estimate. Histograms on the marginal axes show univariate distributions for each variable in the patient group. Corners: Corner boxes show internal representations (top) and behavioral series of responses (bottom) for 4 illustrative patients. Patients in top corners have internal representations (blue) that are similar to controls (orange), but vary in amounts of internal noise (e.g. showing excessive response perseveration; top-right). Patients in bottom corners have atypical representations (blue), but some nevertheless retain healthy levels of internal noise (e.g., being normally consistent in wrongly expecting question phrases to decrease rather than increase in pitch; bottom-left). The estimation of internal noise was limited to the range [0; + 5std]; data points in the upper side of that range may either correspond to true internal noise values, or to larger values for which we could not provide an exact estimate, as illustrated here with a dotted line in the central panel (see Appendix for details).

internal noise) have better sensitivity than MEC for separating high-functioning patients from controls. Finally, both measures appear to have relatively good specificity with respect to non-prosody-related impairments of auditory attention and auditory processing, although internal noise was also found associated with mood disorders which, in our sample, were also predictors of MEC scores.

The fact that abnormal internal representations in our patient sample correlate with performance in prosody recognition but not repetition prompts the question whether impairments in perceptive representations are dissociated from impairments in mapping process between these representations and the corresponding phonatory and articulatory commands involved in their production. On the one hand, the MEC "repetition task", which consists of hearing a target expression produced by the therapist and subsequently reproducing it vocally, does not necessarily involve perceptual representations associated with the recognition of the expression as being e.g. interrogative. It could in principle result from the direct sensorimotor mapping of the auditory characteristics of the stimuli to the corresponding pattern of phonatory (in the case of pitch) and articulatory (in the case e.g. of phonemes or timbre) motor commands. It follows that low recognition scores could in principle be associated with good repetition scores (which indeed we're seeing in a good share of our patients, see upper-left quadrant in Supplementary Fig. S1). This pattern of results is consistent e.g. with literature showing imitation of vocal gestures (such as smiling) without their simultaneous recognition[35]. On the other hand, a wealth of research has documented strong links between action and perception in imitation tasks, and notably established that imitation or action simulation has a causal role in facilitating recognition[36]. For instance, blocking the imitation of a facial expression has detrimental behavioural[37] and neurophysiological effects[38] on their simultaneous recognition. In that sense, it could be expected that patients with low repetition scores would also show low recognition scores (which is again consistent with the low number of data point in the bottom-right quadrant of Fig. S1). To further investigate these links, it would be interesting to collect additional data in which we specifically ask patients to vocalize interrogative prosodies (without providing any auditory examples), and examine the correspondence between their recognition kernel and their produced pitch profiles.

The fact that a majority of patients tested with abnormal melodic processing abilities (MBEA < 65) also had impaired prosodic representations (although the opposite was not true, see Fig. S2) brings questions about the

sensory/cognitive resources shared between speech and music processing. First, this pattern of results suggests that melodic processing and prosodic representations are at least partially subtended by common domain-generic mechanisms, plausibly linked to pitch contour processing. Such a mechanism would be consistent with previous research showing impairments of amusic patients in distinguishing questions from statements[30], emotional prosody[39] and discriminating lexical tones in tone language[40]. Second, it remains that intact pitch/melodic processing is not sufficient to maintain intact prosodic representations (which are impaired in 2/5 of MBEA-positive patients; Fig. S2-top). For this latter subset of patients, impaired representations could result from higher-level lexical or semantic impairments such as difficulties integrating pitch and phonemic information (e.g. failing to associate increasing pitch with the second phoneme of the word "vrai-ment"), or from an impaired semantic representation of what is a question and how it should sound like (e.g. some patients may be consciously expecting that questions are associated with an initial pitch rise). This would be consistent with previous research showing stronger evidence of shared processes between speech and music at earlier and subcortical levels[41] than e.g. in processes of lexical or semantic verification[42]. Further work could look at these possibilities by e.g. testing patients with monosyllabic words (aah?) or a non-semantic task in which patients have to identify which of two alternatives sounds more like a sound target (which only incidentally sounds like a question).

More generally, while our study includes right-hemisphere damage (RHD) patients based on a wealth of clinical literature associating stroke-related RHD with receptive aprosodia[1,4,5,43], our results are only correlational and merely observing changes in internal representation and internal noise in patients with right-hemisphere lesions does not necessarily mean these effects are caused by the right hemisphere damage. Without more direct evidence, one can only speculate about the possible neurological bases for these two types of abnormalities. In terms of representations, one might imagine the involvement of sensory areas, possibly lateralized and specialized for e.g. vocal sounds and/or the slow-varying spectral changes that are characteristics of prosodic pitch contours (e.g. right STG[44]). Regarding internal noise, we may be looking at more diffuse causes, possibly involving frontal areas, and possibly less lateralized[4]. To further look into these questions, future studies could examine possible dissociations with other types of lesions (typically, are left hemisphere stroke patient similarly impaired in representations and/or noise) or use lesion-symptom mapping approaches within a RHD group to link both types of impairment to possibly more specific right areas[46].

In this study, we have focused on a specific type of linguistic prosodic function, namely the marking of interrogation by a final pitch rise. Our focus on interrogative prosody in the present task should by no means be taken as a proposal that it constitutes the optimal test providing most coverage of stroke-related prosodic impairments, but rather as a proof of concept. The reverse-correlation paradigm lends itself ideally to investigate a wide range of other tasks, such as pitch contour representations in other types of linguistic prosody (e.g. imperative sentences to complement the items available in MEC, or prosodic cues to word boundaries[47]), emotional or social prosody (e.g. dominance and trustworthiness[14]); but also other acoustic domains that pitch, such as loudness and speech rate[48] or timbre/phonological cues as used e.g. in phoneme classification[49]. Because of its versatility, reverse correlation appears as a promising way to evaluate prosodic perception mechanisms mechanistically across such a wide range of tasks and cues.

In sum, the representation + noise model paints a simple yet potent portrait of the variety of sensory/cognitive mechanisms that can explain impairments of prosody processing after stroke: patients may differ from controls by having altered representations but a healthy level of internal noise (e.g., being *normally* consistent in *wrongly* expecting e.g. question phrases to decrease rather than increase in pitch—Fig. 3-left); by having normal representations but abnormal levels of internal noise (e.g. showing excessive response perseveration and suboptimal executive control on top of otherwise *normal* sensory processing—Fig. 3-right); or both.

By separating these different profiles of pathology, it is our hope that the representation + noise model will provide more effective and individualized therapeutic targets for rehabilitation of individuals with impaired speech prosody perception than existing measures[50]. Our data indicate that deficits in prosody perception can stem both from attentional/executive or representational problems, underscoring our approach's utility in revealing the underlying mechanisms behind individual patients' comprehension difficulties. Importantly, not all patients with attentional challenges will exhibit aprosodia[1], which positions our method as a complement to, rather than a replacement for, traditional attention assessments by pinpointing the specific contributors to perceptual difficulties. This effort aims to enrich our understanding and assessment of the complex nature of prosody perception and its deficits. For example, patients with the highest levels of internal noise may benefit from therapies that focus on attentional and executive skills, or from transcranial brain stimulation, which has been found to selectively manipulate internal noise in visual tasks[51]. Similarly, for patients encountering difficulties at the internal representation stage, targeted interventions could emphasize pitch contour discrimination or melody imitation tasks, potentially augmented with visual feedback to bolster the reformation of accurate internal representations of prosodic and musical elements[8,52]. Finally, regarding clinical functionality, while the reverse-correlation procedure is, for now, comparable in duration with the MEC perception tasks (MEC: M = 10–15 min, revcor: M = 15 min), it is also easy to dispense remotely (the current control sample was collected with an online app, https://github.com/neuro-team-femto/revcor), does not require supervision or manual scoring, and can be optimized to even shorter durations using e.g. genetic programming optimizations that continuously adapt the presented stimuli to the patient's previous responses[53]. With such adaptations, the reverse correlation procedure could be used to evaluate the prognostic value of measuring changing levels of representation typicality and noise longitudinally, along the weekly or even daily course of rehabilitation.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Code availability

Experimental procedures, stimuli, code for analyzing the data, as well as a selection of control (but not patient) data are available at https://github.com/neuro-team-femto/revcor_avc_public. The online application for collecting control data is available at https://github.com/neuro-team-femto/revcor. Sound stimuli were generated with the CLEESE toolbox, Python language, v1.0, available at https://github.com/neuro-team-femto/cleese. Kernel and internal noise analysis done with the PALIN toolbox, Python language, v1.0, available at https://github.com/neuro-team-femto/palin.

Received: 19 December 2023; Accepted: 6 June 2024
Published online: 02 July 2024

References

1. Blake, M. L., Duffy, J. R., Myers, P. S. & Tompkins, C. A. Prevalence and patterns of right hemisphere cognitive/communicative deficits: Retrospective data from an inpatient rehabilitation unit. *Aphasiology* **16**, 537–547 (2002).
2. Côté, H., Payer, M., Giroux, F. & Joanette, Y. Towards a description of clinical communication impairment profiles following right-hemisphere damage. *Aphasiology* **21**, 739–749 (2007).
3. Villain, M. *et al.* Affective prosody and depression after stroke: A pilot study. *Stroke* **47**, 2397–2400 (2016).
4. Stockbridge, M. D. *et al.* Aprosodia subsequent to right hemisphere brain damage: A systematic review and meta-analysis. *J. Int. Neuropsychol. Soc.* **28**, 709–735 (2022).
5. Tompkins, C. A., Klepousniotou, E. & Scott, A. G. Nature and assessment of right hemisphere disorders. *Aphasia Relat. Neurogenic Commun. Disord.* **2012**, 297–332 (2012).
6. Etchepare, A. & Prouteau, A. Toward a two-dimensional model of social cognition in clinical neuropsychology: A systematic review of factor structure studies. *J. Int. Neuropsychol. Soc.* **24**, 391–404 (2018).
7. Joanette, Y. *et al.* Protocole Montréal d'Evaluation de la Communication (MEC). *Isbergues Fr. Ortho Ed.* **2004**, 896 (2004).
8. Benedetti, V., Weill-Chounlamountry, A., Pradat-Diehl, P. & Villain, M. Assessment tools and rehabilitation treatments for aprosodia following acquired brain injury: A scoping review. *Int. J. Lang. Commun. Disord.* **57**, 474–496 (2022).
9. Schirmer, A. & Kotz, S. A. Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends Cogn. Sci.* **10**, 24–30 (2006).
10. Grandjean, D. Brain networks of emotional prosody processing. *Emot. Rev.* **13**, 34–43 (2021).
11. Zatorre, R. J., Belin, P. & Penhune, V. B. Structure and function of auditory cortex: Music and speech. *Trends Cogn. Sci.* **6**, 37–46 (2002).
12. Read, J. C. A. The place of human psychophysics in modern neuroscience. *Neuroscience* **296**, 116–129 (2015).
13. Murray, R. F. Classification images: A review. *J. Vis.* **11**, 2–2 (2011).
14. Ponsot, E., Burred, J. J., Belin, P. & Aucouturier, J.-J. Cracking the social code of speech prosody using reverse correlation. *Proc. Natl. Acad. Sci.* **115**, 3972–3977 (2018).
15. Neri, P. How inherently noisy is human sensory processing?. *Psychon. Bull. Rev.* **17**, 802–808 (2010).
16. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).
17. Hoyland, A. *et al.* Reverse correlation uncovers more complete tinnitus spectra. *IEEE Open J. Eng. Med. Biol.* **4**, 116–118 (2023).
18. De Boer, E. & De Jongh, H. R. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *J. Acoust. Soc. Am.* **63**, 115–135 (1978).
19. Vilidaite, G., Yu, M. & Baker, D. H. Internal noise estimates correlate with autistic traits: Internal noise correlates with autistic traits. *Autism Res.* **10**, 1384–1391 (2017).
20. Park, W. J., Schauder, K. B., Zhang, R., Bennetto, L. & Tadin, D. High internal noise and poor external noise filtering characterize perception in autism spectrum disorder. *Sci. Rep.* **7**, 17584 (2017).
21. Wang, L. *et al.* Mental representations of speech and musical pitch contours reveal a diversity of profiles in autism spectrum disorder. *Autism* **2022**, 27. https://doi.org/10.1177/13623613221111207 (2022).
22. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15 (2002).
23. Ukaegbe, O. C. *et al.* Aprosodia following focal brain damage: What's right and what's left?. *Am. J. Speech Lang. Pathol.* **31**, 2313–2328 (2022).
24. Le Lafon, J. C. *Test Phonétique et la Mesure de l'Audition* (Springer, 1964).
25. Mazaux, J. M. & Orgogozo, J. M. Boston Diagnostic Aphasia Examination (BDAE-F). *Issy-Moulineaux Fr. Ed. Sci. Psychol.* **1981**, 526 (1981).
26. Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta Psychiatr. Scand.* **67**, 361–370 (1983).
27. Ambert-Dahan, E. *et al.* Capacités attentionnelles auditives et presbyacousie. *Rev. Neurol. (Paris)* **169**, A236–A237 (2013).
28. Tessier, C. & Weill-Chounlamountry, A. Aide informatisée pour la rééducation des troubles auditifs centraux 2 (Airtac2). *Paris Fr. Ortho Ed.* **2014**, 236 (2014).
29. Peretz, I., Champod, A. S. & Hyde, K. Varieties of musical disorders : The Montreal battery of evaluation of Amusia. *Ann. N. Y. Acad. Sci.* **999**(1), 58–75 (2003).
30. Patel, A. D., Wong, M., Foxton, J., Lochy, A. & Peretz, I. Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Percept.* **25**, 357–368 (2008).
31. Burred, J. J., Ponsot, E., Goupil, L., Liuni, M. & Aucouturier, J.-J. CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PLOS ONE* **14**, e0205943 (2019).
32. Burgess, A. E. & Colborne, B. Visual signal detection IV Observer inconsistency. *J. Opt. Soc. Am. A* **5**, 617 (1988).
33. Nikitina, L., Paidi, R. & Furuoka, F. Using bootstrapped quantile regression analysis for small sample research in applied linguistics: Some methodological considerations. *PLOS ONE* **14**, e0210668 (2019).
34. Eshin, J. Pymer4: Connecting R and Python for linear mixed modeling. *J. Open Sourc. Softw.* **3**(31), 862 (2018).
35. Arias, P., Belin, P. & Aucouturier, J.-J. Auditory smiles trigger unconscious facial imitation. *Curr. Biol.* **28**, R782–R783 (2018).
36. Blakemore, S.-J. & Decety, J. From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* **2**, 561–567 (2001).
37. Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J. & Davidson, R. J. Cosmetic use of botulinum toxin-a affects processing of emotional language. *Psychol. Sci.* **21**, 895–900 (2010).
38. Stark, S., Stark, C., Wong, B. & Brin, M. F. Modulation of amygdala activity for emotional faces due to botulinum toxin type A injections that prevent frowning. *Sci. Rep.* **13**, 3333 (2023).
39. Thompson, W. F., Marin, M. M. & Stewart, L. Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proc. Natl. Acad. Sci.* **109**, 19027–19032 (2012).
40. Nguyen, S., Tillmann, B., Gosselin, N. & Peretz, I. Tonal language processing in congenital amusia. *Ann. N. Y. Acad. Sci.* **1169**, 490–493 (2009).

41. Bidelman, G. M., Gandour, J. T. & Krishnan, A. Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain Cogn.* **77**, 1–10 (2011).
42. Sadakata, M., Weidema, J. L. & Honing, H. Parallel pitch processing in speech and melody: A study of the interference of musical melody on lexical pitch perception in speakers of Mandarin. *PLOS ONE* **15**, e0229109 (2020).
43. Sheppard, S. M. *et al.* The company prosodic deficits keep following right hemisphere stroke: A systematic review. *J. Int. Neuropsychol. Soc.* **28**, 1075–1090 (2022).
44. Tang, C., Hamilton, L. S. & Chang, E. F. Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797–801 (2017).
45. Vilidaite, G., Marsh, E. & Baker, D. H. Internal noise in contrast discrimination propagates forwards from early visual cortex. *NeuroImage* **191**, 503–517 (2019).
46. Sheppard, S. M. *et al.* Characterizing subtypes and neural correlates of receptive aprosodia in acute right hemisphere stroke. *Cortex* **141**, 36–54 (2021).
47. Osses, A., Spinelli, E., Meunier, F., Gaudrain, E. & Varnet, L. Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation. *JASA Express Lett.* **3**, 095205 (2023).
48. Goupil, L., Ponsot, E., Richardson, D., Reyes, G. & Aucouturier, J.-J. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nat. Commun.* **12**, 861 (2021).
49. Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F. & Hoen, M. A psychophysical imaging method evidencing auditory cue extraction during speech perception: A group analysis of auditory classification images. *PLOS ONE* **10**, e0118009 (2015).
50. Tompkins, C. A. Rehabilitation for cognitive-communication disorders in right hemisphere brain damage. *Arch. Phys. Med. Rehabil.* **93**, S61–S69 (2012).
51. Vilidaite, G. & Baker, D. H. Psychophysical measurement of the effects and non-effects of TMS on contrast perception. *Brain Stimul.* **11**, 956–957 (2018).
52. Rosenbek, J. C. *et al.* Novel treatments for expressive aprosodia: A phase I investigation of cognitive linguistic and imitative interventions. *J. Int. Neuropsychol. Soc.* **10**, 786–793 (2004).
53. Binetti, N. *et al.* Genetic algorithms reveal profound individual differences in emotion recognition. *Proc. Natl. Acad. Sci.* **119**, e2201380119 (2022).

Author contributions

M.J., J.-J.A., M.V., E.P. and L.N. contributed to the conception and design of the study. All authors contributed to the acquisition and analysis of data. A.A.Z., J.-J.A., M.V., M.J. and E.P. contributed to drafting the manuscript and preparing the figures. All authors reviewed the manuscript.

Funding

This work was supported by grant from *Fondation pour l'Audition* (FPA RD 2021-12).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-64295-y.

Correspondence and requests for materials should be addressed to M.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.3 Conclusion: Revcor parameters as potential biomarkers

In this study, we focused on a specific aspect of linguistic prosody: the marking of interrogation through a final pitch rise. Our choice of interrogative prosody serves as a proof of concept rather than a comprehensive test of stroke-related prosodic impairments. The reverse correlation paradigm is highly adaptable and could be applied to various prosodic functions, such as imperative sentences, emotional prosody (e.g., dominance, trustworthiness) Ponsot, Burred, et al. (2018), or other acoustic features like loudness, speech rate Goupil, Ponsot, Richardson, Reyes, and Aucouturier (2021b), and phonological cues Osses, Spinelli, Meunier, Gaudrain, and Varnet (2023). Its versatility makes it a valuable tool for mechanistically assessing prosodic perception across different contexts.

Ultimately, representation + noise model offers a framework for understanding prosody impairments after stroke, distinguishing between deficits in internal representations, abnormal internal noise, or a combination of both.

CHAPTER 5

Limitations of the classical estimation methods: a posthoc analysis of accuracy

In the preliminary study presented in Chapter 4, analyzing the patients' kernels and internal noise suggested a way to identify distinct pathological profiles and uncover the diverse sensory and cognitive mechanisms underlying prosody processing impairments after stroke. These two biomarkers hold significant potential for clinical applications, making their precise estimation crucial to ensure their reliability as long-term diagnostic tools. Also, since not all patients exhibited deficits in both parameters, examining extreme cases can provide valuable insights into the potential neurological bases of these abnormalities. However, an important question remains: are these cases truly extreme, or do the classical estimation methods used in Chapter 4 (namely: kernel: classification images; noise: double pass) fail to accurately capture their impairments? Various factors could contribute to the uncertainty in these estimates. Some of them are already discussed in the reverse correlation literature, for instance the number of trials used in the experiment, and others specific to our patient population, namely their tendency to perseverate in their responses. We present here a post-hoc analysis of estimation accuracy

with these two factors, and conclude on how they impact the quality of estimates. This chapter concludes Part II and draw a roadmap for the methodological contributions in Part III.

5.1 Number of trials

In a reverse correlation experiment, one fundamental challenge is determining the sufficient number of trials required to obtain a robust estimation of both kernel and internal noise. Reverse correlation relies on stimulus sampling, meaning that more trials generally provide better statistical power. However, in a clinical population, practical constraints such as cognitive fatigue and attention skills must also be considered when setting trial numbers.

5.1.1 Number of trials as a challenge to stroke patients

Stroke patients often experience post-stroke fatigue (Paciaroni & Acciarresi, 2019), which can significantly impact their ability to maintain focus and perform demanding tasks over time (Cumming et al., 2016). In addition, selective attention (as discussed in Chapter 2), a critical capacity for filtering relevant information and ignoring distractions, can be particularly vulnerable following stroke. Both fatigue and attention deficits raise important concerns about the number of experimental trials stroke patients can realistically complete, as excessive cognitive strain may compromise their performance and the validity of the data collected.

In our study, we limited the number of trials to 150 (100+50), with an additional 50 double-pass trials for estimating internal noise. While this number is already much smaller than typical experiments (700 in Ponsot, Burred, et al. (2018), several 1000s in Neri (2010)), it is an open question whether this number of trials was already inducing fatigue in patients. To investigate this, we analyzed their response times across the 150 trials (see Figure 5.1-top). As expected, post-stroke patients demonstrated generally slower response times on cognitive tasks compared to controls (Patients, $m = 1.91$ s; Controls,

$m = 1.36$ s) (Cumming, Brodtmann, Darby, & Bernhardt, 2012; Feigin et al., 2010), consistent with the well-documented slowing of cognitive processing following stroke. Our results did not reveal a consistent increase with trial number, which would typically be expected if fatigue played a dominant role (Goh & Stewart, 2019). However, we observed that in RHD patients, response times increased toward the end of Block 1, potentially reflecting transient fatigue or attention fluctuations (see Figure 5.1-top). Interestingly, response times decreased after the break before Block 2, which we speculate may reflect a combination of recovery from fatigue and task familiarization or learning effects.

These are observational trends; to formally assess RT dynamics across the entire session, we conducted a linear mixed-effects analysis on log-transformed response times. The model included fixed effects for group (patients vs. controls), trial number, and their interaction, with a random intercept for participant. Results showed that patients were significantly slower overall ($\beta = 0.234$, $p = 0.032$), but both groups demonstrated a significant decrease in response times across trials ($\beta = -0.002$, $p < 0.001$). The interaction term between groups and trial number was also significant but small ($\beta = -0.0004$, $p = 0.002$), indicating a slightly steeper reduction in response times for patients over time.

It is important to note, however, that changes in response time do not directly map onto performance or fatigue per se. Several interpretations are possible: for instance, patients might respond more quickly as the session progresses simply to "get it over with," rather than due to increased familiarity or reduced cognitive effort. Conversely, longer response times could reflect greater engagement and a more deliberate attempt to answer correctly, rather than cognitive slowing or fatigue. Thus, the relationship between reaction times, task engagement, and cognitive fatigue remains complex and multifactorial. Previous work has shown that stroke survivors experience within-block fluctuations in task engagement, sometimes alleviated by breaks, but these temporal dynamics cannot be solely attributed to a simple cumulative fatigue effect (Brosnan et al., 2020).

Furthermore, when examining average response times across three separate blocks of 50 trials each, we observed a noticeable decrease in response times between the first and second blocks (see Figure 5.1-bottom left). A linear mixed-effects model on log-

transformed response times revealed a significant main effect of block ($\beta = -0.104$, $p < 0.001$), indicating a systematic reduction in response time across blocks. The model also showed a significant group-by-block interaction ($\beta = -0.026$, $p < 0.001$), suggesting that patients exhibited an even greater decrease in response time across blocks than controls. Controls' average response times decreased from 1.63 seconds in Block 1 to 1.25 in Block 2 and 1.21 in Block 3. In patients, the drop was more pronounced, from 2.41 seconds in Block 1 to 1.68 in Block 2 and 1.66 in Block 3. This pattern suggests two possible interpretations for controls and patients: a learning effect, in which faster response times reflect improved task performance after initial exposure to the task structure; or loss of attention, in which patients become less engaged in the task after the first block, leading to faster but potentially less controlled responses.

Examining kernel typicality (the similarity between a kernel and the kernel of a control group), healthy participants show a progressive improvement in the typicality of their mental representation across consecutive blocks (mean typicality: 0.84 in Block 1, 0.89 in Block 2, and 0.89 in Block 3). In contrast, stroke patients exhibit a decline in kernel typicality across blocks (mean typicality: 0.58 in Block 1, 0.56 in Block 2, and 0.50 in Block 3; see Figure 5.1-bottom right). A linear mixed-effects model confirmed this pattern, revealing a significant group-by-block interaction ($\beta = -0.064$, $p = 0.033$), indicating that typicality increased across blocks in controls but decreased in patients. This divergence suggests distinct learning trajectories between groups, potentially reflecting impaired task-related representation or retention mechanisms in stroke patients.

5.1.2 Effect of number of trials on kernel estimation

While the reverse correlation literature often discusses the choice of number of trials as affecting the accuracy of kernel estimation (e.g., Burred et al. (2018)), precise simulations of that effect are rarely reported. In particular, it appears plausible that, at low number of trials, the effect of larger internal noise (as seen in our patients; see Chapter 4) has a large influence on kernel estimation.

To benchmark the accuracy of kernel estimation that's achievable with 150 trials, we

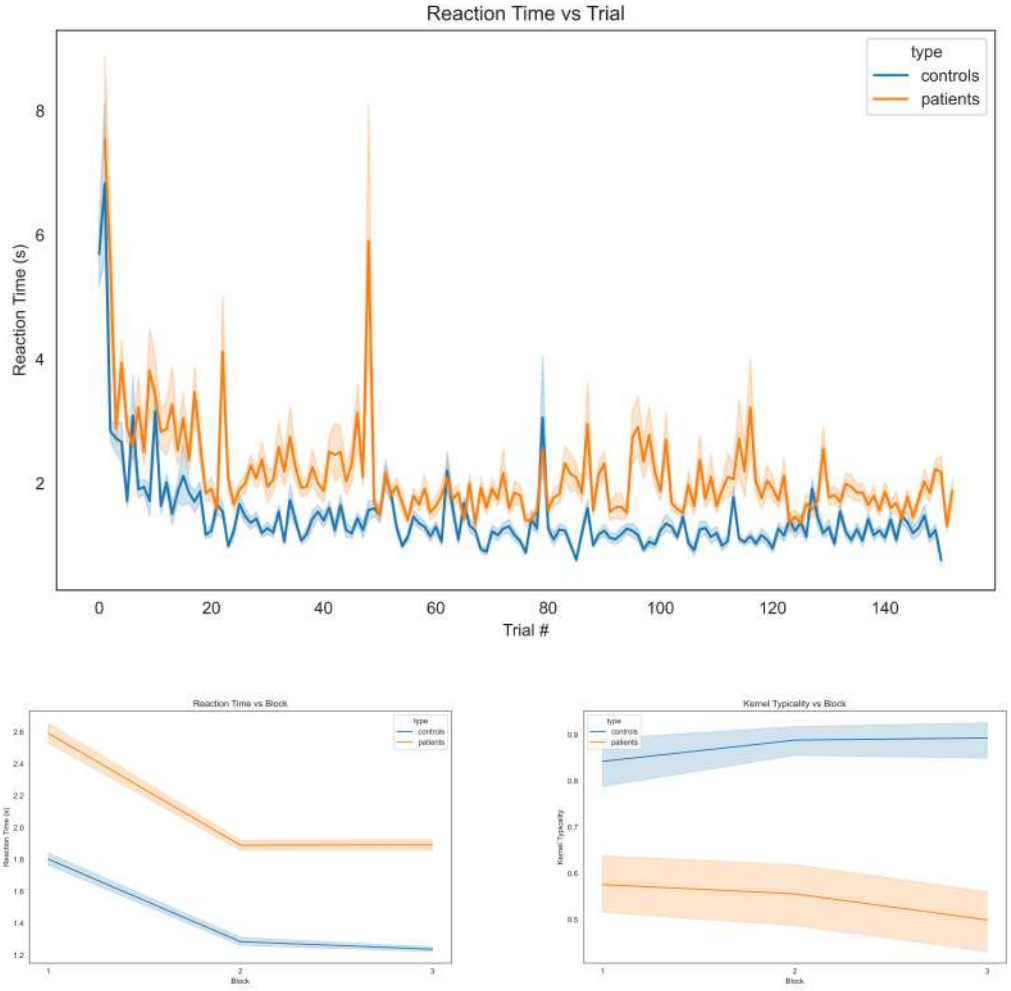


Fig. 5.1 **Reaction times and kernel typicality throughout the experiment Top:** Mean reaction times across trials for stroke patients (orange) and healthy controls (blue). While patients show generally slower RTs throughout, both groups exhibit a steep initial drop suggesting early task familiarization. **Bottom left:** Mean RTs across the three blocks for stroke patients and controls. Patients exhibit a marked decrease in RT from Block 1 to Block 2, suggesting a possible recovery from fatigue or increased task familiarity after the first break. Control participants show minimal change across blocks. **Bottom right:** Kernel typicality by block. Controls improved across blocks, while patients showed declining typicality, suggesting increased noise or perseveration.

used the PALIN toolbox. PALIN enables the simulation of a variety of linear observers performing the same reverse correlation experiment as real participants, while we control their true level of internal noise and criteria. We simulated an experiment with the same number of trials as our study (n=150). The observer is assigned a known true kernel

and completes 50 double-pass trials. This simulation is repeated 1000 times to estimate the mental representation using two methods introduced in Chapter 3: the weighted-average ("classification image") method and the GLM method. Estimation error on the kernel is evaluated by calculating the Pearson correlation between the true kernel and the estimated kernel at different levels of true internal noise.

Palin Toolbox: Impact of internal noise on kernel recovery with limited trials

This simulation explores how varying internal noise levels (from 0 to 5) affect kernel recovery accuracy using 150 trials per observer. Two estimation methods (classification image and GLM) are compared across 1000 runs per condition, using kernel correlation as the similarity metric.

```
observer_params = {'kernel':['random'],
                   'internal_noise_std':np.arange(0,5.1,0.5),
                  'criteria':[0]}
experiment_params = {'n_trials': [150], 'trial_type': [Int2Trial],
    'n_features': [6],'external_noise_std': [100] }
analyser_params = {'kernel_extractor': [ClassificationImage,GLMKernel],
                    'distance': ['CORR']}
sim_kernel = Sim(SimpleExperiment, experiment_params,
                 LinearObserver, observer_params,
                 KernelDistance, analyser_params)
sim_kernel_fix = sim_kernel.run_all(n_runs=1000)
```

Results are shown in Figure 5.2. Both kernel estimation methods demonstrate high kernel correlation at lower internal noise levels, with higher precision for the GLM method, but a progressive decline is observed as internal noise increases. At internal noise levels representative of healthy participants ($M\sigma_n = .8$, in Chapter 4), both methods reach correlations > 0.95. However, at levels representative of patients ($M\sigma_n = 2.8$, in Chapter 4), the correlation degrades to 0.92.

In general, the limited number of trials ($n = 150$) constrains the precision of kernel estimation, especially at higher noise levels. This limitation is evident in the confidence intervals, which reflects greater variability in estimation as internal noise increases.

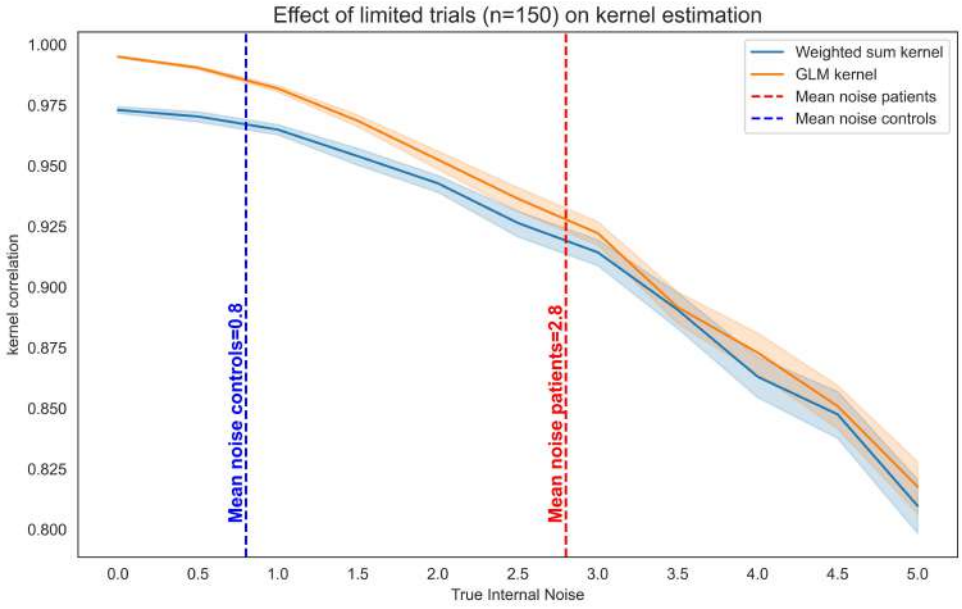


Fig. 5.2 **Impact of internal noise on kernel recovery with limited trials ($n = 150$)** Effect of limited trials ($n = 150$) on kernel estimation accuracy (y-axis) as a function of true internal noise (x-axis). Two estimation methods are compared: weighted sum kernel (orange) and GLM kernel (blue). Kernel accuracy decreases with increasing noise for both methods. Vertical dashed lines indicate the mean estimated internal noise for controls (blue, $\sigma = 0.8$) and stroke patients (red, $\sigma = 2.8$). This highlights the greater impact of internal noise on kernel estimation quality in the patient group.

5.1.3 Effect of number of trials on internal noise estimation

Similarly, the small number of trials, and in particular the number of repeated double-pass trials (2x $n = 50$), is likely to have an important impact on the accuracy of internal noise estimation. Again, while this factor is sometimes discussed in the reverse correlation literature (incl. in our own work Adl Zarrabi et al. (2024)), the impact is rarely quantified and the typical solution for healthy participants is simply to increase the number of double-pass trials.

To benchmark the accuracy of internal noise estimation achieved with the $n = 150$ trials of Chapter 4, we used PALIN to simulate an experiment with 150 trials, including 50 double-pass trials, repeated 1000 times to evaluate the estimation error of internal noise over a range of linear observers with known levels of true internal noise between $[0,5]$. The percentage error is calculated as the absolute difference between the true and

estimated internal noise, normalized by the true value, and expressed as a percentage.

Palin Toolbox: Internal noise estimation error using Double-pass method with limited trials

This simulation evaluates internal noise estimation using the double-pass method, where 50 of 150 trials are repeated. Internal noise values are systematically varied between 0 and 5 (in steps of 0.1), and estimation accuracy is assessed over 1000 simulated observers using a precomputed agreement model (lookup 3.4).

```
observer_params = {'kernel':['random'],
                   'internal_noise_std':np.arange(0,5.1,0.1),
                  'criteria':[0]}
experiment_params = {'n_trials':[150], 'n_repeated':[50],
                     'trial_type': [Int2Trial],'n_features': [6],
                     'external_noise_std': [100]}
analyser_params = {'internal_noise_extractor':[DoublePass],
                   'agreement_model_file':['agreement_model_large.csv']}
sim_in = Sim(DoublePassExperiment, experiment_params,
          LinearObserver, observer_params,
          InternalNoiseValue, analyser_params)
sim_in_fix = sim_in.run_all(n_runs=1000)
```

The estimation error with the double-pass method with 50 repeated trials achieved an unimpressive 40% relative error for noise levels between [0.5, 3.5] and increased to 50% for noise levels exceeding 3.5 (3.5). The error was comparable at internal noise levels representative of controls ($M\sigma_n = .8$) and patients ($M\sigma_n = 2.8$), although it was – perhaps counter-intuitively – a few percent points greater for controls.

While estimation error was null at zero internal noise (corresponding to $p_{agree} = 1$), the error was very large at small internal noise levels ($[0, 0.5]$. This is both due to small-number sampling errors and to the fixed grid size used for double-pass lookup (see Chapter 3 for details). First, small internal noise corresponds to very low probabilities of disagreement which, sampled at only $n = 50$ repeated trials, may across the experiment correspond to only 1 or 2 trials with inconsistent responses. Second, because we used a lookup table with a 0.2 search grid on internal noise, such small fluctuations may create large % errors. This situation could be improved with lookup tables with non-linear grid resolution (e.g.,

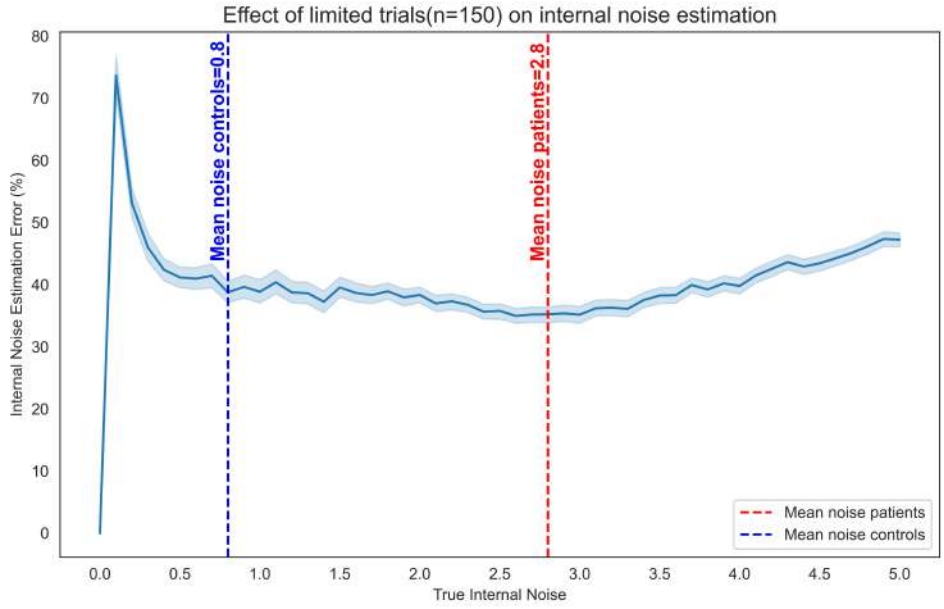


Fig. 5.3 **Internal noise estimation error using Double-pass method ($n = 150$)** Effect of limited trials ($n = 150$) on internal noise estimation using the double-pass method. The y-axis shows the percentage error in estimated internal noise as a function of the true internal noise. Estimation error remains on average around 40% and relatively stable across a wide range of internal noise levels but increases at very low and high extremes. Vertical dashed lines mark the average internal noise levels for stroke patients (red, $\sigma = 2.8$) and healthy controls (blue, $\sigma = 0.8$), demonstrating that double-pass estimates are most accurate within this central range.

logarithmic), or with optimisation methods that alleviate the need for a fixed search grid. It is to be noted that the estimation of low internal noise levels in the range $[0, 0.5]$ is almost only a theoretical problem, as typical values for healthy controls are more typically around 1. We will return to the problem of small internal noise estimation in Chapter 6.

In summary, while the clinical constraints of deploying reverse correlation during speech therapy sessions compel us to use only a limited number of experimental trials (fixed here at $n = 150$, incl. 2x50 double-pass), a post-hoc analysis with simulations shows that this setup in fact severely limits the accuracy of estimates. While kernel correlation degrades with increasing patient internal noise, it remains at an acceptable 0.90 (Figure 5.2). However, double-pass estimation of internal noise estimation at n=150 severely suffers at about $40 - 50\%$ error (Figure 5.3), casting serious doubts on internal noise results obtained in Chapter 4. While the quality of internal noise estimation would

improve by increasing the number of trials (Chapter 3 - Figure 3.5, see also Chapter 6 below), the analysis of kernel correlation across blocks (Figure 5.1) suggests that patients, contrary to controls, are already at a limit in terms of experiment duration, and that increasing the number of trials is impractical.

5.1.4 Perseveration phases observed

By examining our patients' responses, we observe that some stroke patients consistently chose the same stimulus across multiple successive trials. Figure 5.4 shows several examples of such patterns, which may be distributed throughout the experiment (subject 29) or localized (subject 40); and become less (subject 41) or more frequent (subject 43) as the experiment progresses.

From the perspective of the linear-observer model, participants exhibiting perseverative behavior no longer rely on their decision model to guide their responses. Instead of making choices based on their kernel, they disengage from actively applying it to the stimuli. It therefore appears suboptimal to evaluate decision parameters (kernel and noise) on the basis of trials for which that decision criteria was not, or only partly applied.

Such perseverative episodes are likely manifestations of underlying attention deficits, particularly in the domain of divided or alternating attention (Cramer et al., 2023). When cognitive resources are insufficient to manage the demands of the task, patients may revert to repetitive, habitual responses rather than flexible, adaptive choices. This view is supported by evidence linking divided attention impairments to behavioral rigidity in neuropsychological populations. Identifying these patterns is essential, as they mark intervals when participants are no longer engaging in goal-directed decision-making, highlighting the need for careful interpretation or exclusion of these trials from analyses focused on perceptual or decision parameters.

In our experiment, we did not have the opportunity to directly measure participants' attention during the blocks. However, one simple way to estimate perseveration is by identifying streaks of repeated responses within each participant, session, and block. By convention, trials are marked as perseverative if they belong to a streak of 15 or more

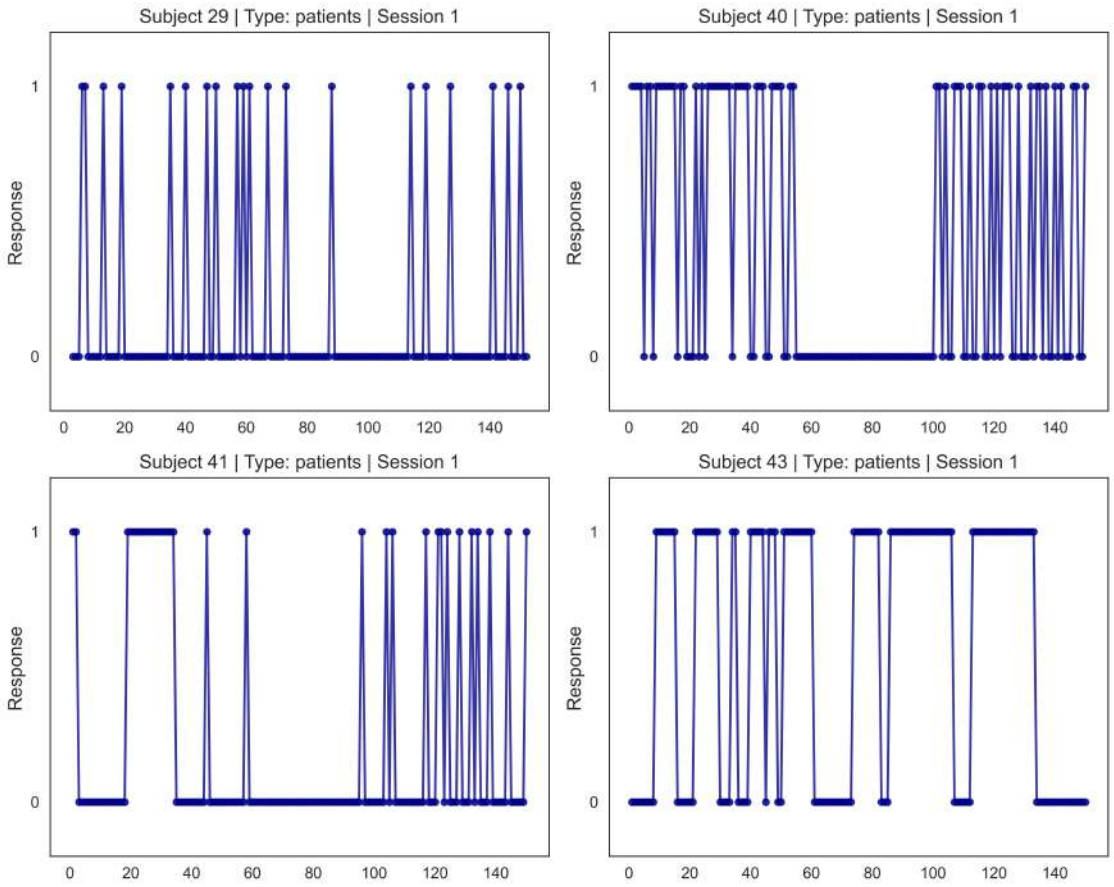


Fig. 5.4 **Variability in perseverative behavior of patients during the Task** Examples of perseverative response patterns in stroke patients. Each panel shows binary responses (0 or 1) over 150 trials for four patients during Session 1. Clear perseverative patterns are observed in all subjects, characterized by long sequences of repeated responses, either sustained 0sec, 1sec. These perseverations can manifest in different temporal patterns which suggest dynamic fluctuations in mental state and task engagement.

consecutive repeated responses, and we compute the "perseveration ratio" as the probability that a trial (out of 150) belongs to such a streak. As shown in Figure 5.5, the perseveration ratio for patients ($median = 0.64$) is higher than controls ($median = 0.56$), particularly in the first and second blocks.

We chose a relatively high threshold (15) for consecutive repeated responses for several reasons. With random stimuli, short runs of repeated responses, such as three or four in a row, can easily occur simply because the presented stimuli happen to align with a participant's internal kernel. In such cases, repeated choices may be entirely appropriate and not reflect true perseverative behavior. Without access to ground-truth labels for

the "correct" response (given the randomness of the stimuli) or an adaptive model that can disambiguate between response alignment and habitual repetition, it is challenging to distinguish genuine perseveration from correct or noisy responses. By setting a higher threshold, we aimed to minimize false positives and focus on clear-cut cases where the likelihood of purely stimulus-driven repetition is low. In Chapter 7 below, we provide a more accurate method to estimate whether trials are perseverated or not.

Palin Toolbox: Perseverating observer

PALIN code for simulating perseverating observers, and compute kernel and internal noise using the Classification Image and GLM Methods (kernels) and the Double Pass method (noise).

```
# Observer parameters define kernel type, internal noise range, decision criteria,
# and state transition probabilities (for perseveration).
observer_params = {'kernel': ['random'],  'internal_noise_std': np.arange(1,5,0.5), 'criteria': [0],
    'transition_matrix': [[[0.9, 0.1], [beta, 1 - beta]] for beta in beta_values]}
# Experiment parameters include trial counts, double-pass trials, stimulus features,
# and external noise level.
experiment_params = {'n_trials':[150], 'n_repeated':[50], 'trial_type': [Int2Trial],
                     'n_features': [7],'external_noise_std': [100]}
# Analyzer parameters for internal noise use the Double Pass method.
analyser_params_noise = {'internal_noise_extractor':[DoublePass],
                   'agreement_model_file':['agreement_model_large.csv']}
# Analyzer parameters for kernel estimation use ClassificationImage and GLMKernel methods.
analyser_params_kernel = {'kernel_extractor': [ClassificationImage,GLMKernel],'distance': ['CORR']}
# Simulate internal noise estimation for perseverating observers.
sim_in_per = Sim(DoublePassExperiment, experiment_params,  PerseveratingObserver, observer_params,
                 InternalNoiseValue, analyser_params_noise)
# Simulate kernel estimation for perseverating observers.
sim_kernel_per = Sim(DoublePassExperiment, experiment_params,  PerseveratingObserver, observer_params,
                 KernelDistance, analyser_params_kernel)
sim_kernel_perseveration_df = sim_kernel_per.run_all(n_runs=1000)
sim_in_perseveration_df = sim_in_per.run_all(n_runs=1000)
```

However, we acknowledge that this operational definition has limitations: a long sequence of repeated responses could occasionally reflect true alignment between stimulus and kernel, and conversely, shorter streaks could sometimes indicate perseveration if they occur in situations where the stimuli are highly contrasted. Thus, this metric may under-

estimate or overestimate the true extent of perseverative episodes.

It is also noteworthy that, even with this strict definition, control participants sometimes exhibited non-negligible perseveration ratios. This could reflect the imprecision of our metric, the presence of similar cognitive fatigue or attentional lapses in healthy controls. We will further explore the implications and limitations of this approach in Chapter 9.

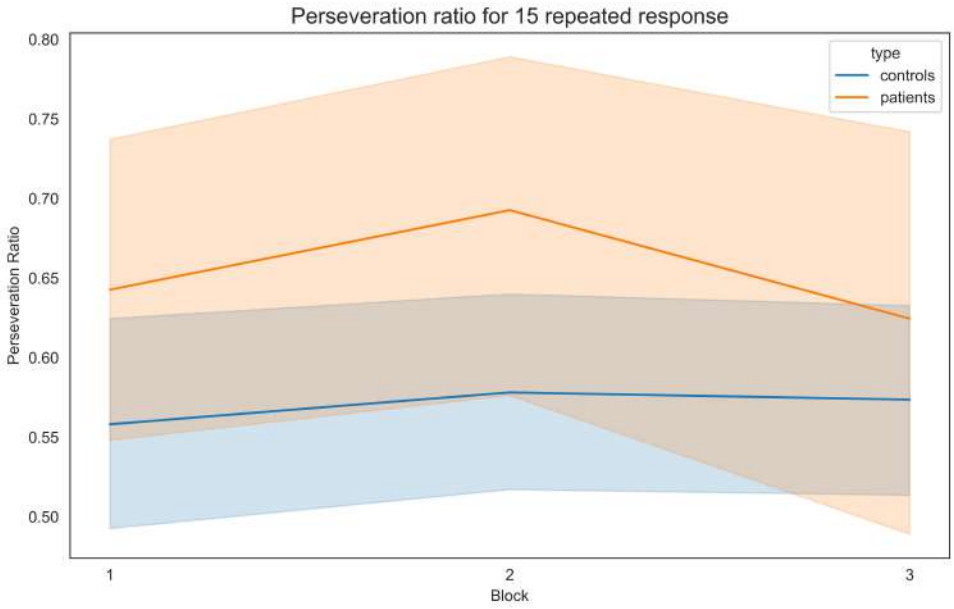


Fig. 5.5 **Perseveration Ratio Across Blocks (≥15 Repeated Responses)** Perseveration ratio across blocks for patients and controls, computed as the proportion of trials belonging to a streak of at least 15 consecutive identical responses. Patients exhibit consistently higher perseveration ratios than controls across all blocks, with a peak in Block 2 followed by a slight decrease in Block 3. In contrast, controls show relatively stable and lower perseveration levels. Shaded areas represent standard errors across participants. These results suggest greater response repetition and reduced flexibility in patients, especially during the middle phase of the task.

5.1.5 Effect of perseveration on kernel estimation

Because kernel estimation relies, roughly speaking, on the average of responses, it is likely that perseveration has an impact on the accuracy of kernel estimation. For instance,

some patients may appear to have representations similar to controls, despite underlying perseverative tendencies.

Using the PALIN framework (see Box 5.1.4), we simulated a perseverating observer with a probability $p(a)$ of entering a perseverative phase and a probability $p(1-b)$ of remaining in that phase. These parameters were varied to increase the proportion of perseverative trials, simulating conditions where patients exhibit more repetitive behavior. In this simulation, the perseverating observer performs the same reverse correlation experiment as real participants, providing insights into how varying levels of perseveration (within a limited number of trials, n=150) influence the accuracy of kernel estimation and, below, that of internal noise. As above, kernel estimation accuracy is estimated by kernel correlation, i.e., the similarity between the true kernel of the observer and the estimated kernel. Like other simulations, this was repeated 1000 times to evaluate the estimation error. In this simulation, the internal noise of the observers was systematically varied between 0 and 5 in order to assess how estimation performance changes across a broad range of noise levels.

Palin Toolbox: Effect of perseverative responding on kernel estimation accuracy with limited trials

This simulation explores how increasing perseveration probability (β) affects kernel estimation accuracy in a two-state observer model. The transition matrix controls the likelihood of remaining in the perseverative state, and kernel recovery is evaluated over 1000 runs using both classification image and GLM methods.

```
beta_values = np.round(np.linspace(0.05, 0.95, 10), 2)
observer_params = {'kernel': ['random'],
    'internal_noise_std': np.arange(1,5,0.5), 'criteria': [0],
'transition_matrix': [[[0.9, 0.1], [beta, 1 - beta]] for beta in beta_values]}
experiment_params = {'n_trials': [150],'n_repeated':[50],
    'trial_type': [Int2Trial],  'n_features': [7],
    'external_noise_std': [100]}
analyser_params = { 'kernel_extractor': [ClassificationImage,GLMKernel],
    'distance': ['CORR'] }
sim = Sim(DoublePassExperiment, experiment_params,
                PerseveratingObserver, observer_params,
                KernelDistance, analyser_params)
sim_df = sim.run_all(n_runs=1000)
sim_df['beta'] = sim_df['transition_matrix'].apply(lambda x: x[1][1])
```

As shown in Figure 5.6, increasing the probability of remaining in the perseverative phase, $p(1-b)$ leads to a noticeable decline in kernel correlation, particularly for probabilities exceeding 0.5. At levels representative of our patient group ($Mp = 0.64$), kernel correlation is about 0.70, a significant decrease from the non-perseverating case of Figure 5.2. This confirms that high levels of perseverative behaviour can distort the relationship between true and estimated kernels.

5.1.6 Effect of perseveration on internal noise estimation

As for kernels, perseveration has the potential to impact the accuracy of internal noise estimation. In double-pass experiments, perseverating behaviour may occur in repeated trials, this may lead either to artificially large probabilities of agreement (repeated the same response throughout the two repeated blocks) or to smaller probabilities when only one block is contaminated. Conversely, if these behaviors do not manifest within the

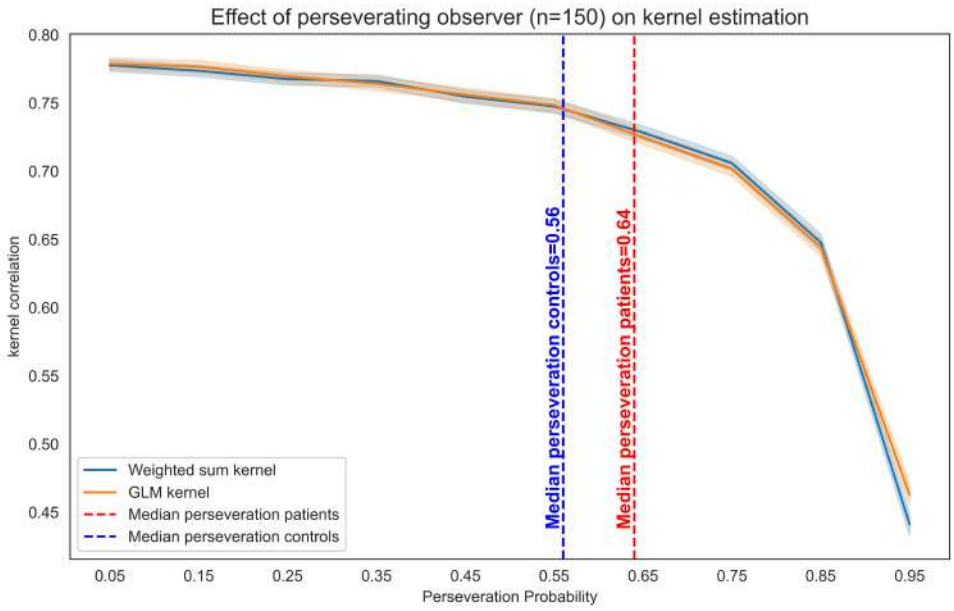


Fig. 5.6 **Effect of perseverative responding on kernel estimation accuracy ($n = 150$)** Kernel estimation accuracy as a function of perseverative behavior, evaluated using simulated observers over 150 trials. Kernel correlation (y-axis) decreases as the probability of perseverative responses increases (x-axis), for both weighted sum kernel (blue) and GLM kernel (orange). The sharp decline above a perseveration probability of 0.75 indicates a critical threshold where estimation quality degrades substantially. Vertical dashed lines show the median perseveration probability for stroke patients (red, 0.64) and controls (blue, 0.56), illustrating that patients tend to operate in a regime with greater impact on kernel reliability.

repeated trials, they may go undetected by internal noise estimates and give the false impression that a patient has normal consistency while they fail to do the task a significant portion of the time. It is in fact an open question whether perseveration should be considered a legitimate part of internal noise or whether it should be estimated separately, leading to situations where patients may display important perseveration but low internal noise anywhere else. Chapter 7 below will propose a method to separate both measures.

Palin Toolbox: Effect of perseverative responding on internal noise estimation by Double-pass

This simulation examines how varying levels of perseveration (β) affect internal noise estimation using the double-pass method.

```
observer_params = {'kernel': ['random'],
    'internal_noise_std': np.arange(1,5,0.5), 'criteria': [0],
    'transition_matrix': [[[0.9, 0.1], [beta, 1 - beta]] for beta in beta_values]}
experiment_params = {'n_trials':[150], 'n_repeated':[50], 'trial_type': [Int2Trial],
                     'n_features': [7],'external_noise_std': [100]}
analyser_params = {'internal_noise_extractor':[DoublePass],
                   'agreement_model_file':['agreement_model_large.csv']}
sim_in_perseveration = Sim(DoublePassExperiment, experiment_params,
          PerseveratingObserver, observer_params,
          InternalNoiseValue, analyser_params)
sim_in_perseveration_df = sim_in_perseveration.run_all(n_runs=1000)
sim_in_perseveration_df['beta'] = sim_in_perseveration_df['transition_matrix'].apply(lambda x: x[1][1])
```

To benchmark the impact of perseveration on internal noise estimation, we used PALIN to create simulations in which perseverating observers with a range of internal noise encounter simulated double-pass experiments, and we measured relative % error in internal noise estimation with the double-pass method. As shown in Figure 5.7, an observer with no perseveration exhibits an estimation error of approximately 40%, consistent with previous findings in Figures 5.3 and 3.5. However, as the probability of perseveration increases, the estimation error initially rises exponentially, with a sharp increase beyond a threshold of perseveration probability $p = 0.55$. Notably, our patient group falls beyond this limit, indicating that higher perseveration levels contribute to significantly greater internal noise estimation errors. This suggests that perseveration disrupts the reliability of internal noise estimation, making it more challenging to accurately capture perceptual variability in perseverating individuals.

Beyond relative error averaged over a range of internal noises, Figure 5.8 shows the impact of perseveration probability β on estimated noise at various noise levels. At low levels of internal noise, perseverating observers exhibit consistent overestimation of noise, and underestimation at larger noise values. Higher probabilities of perseveration (β) am-

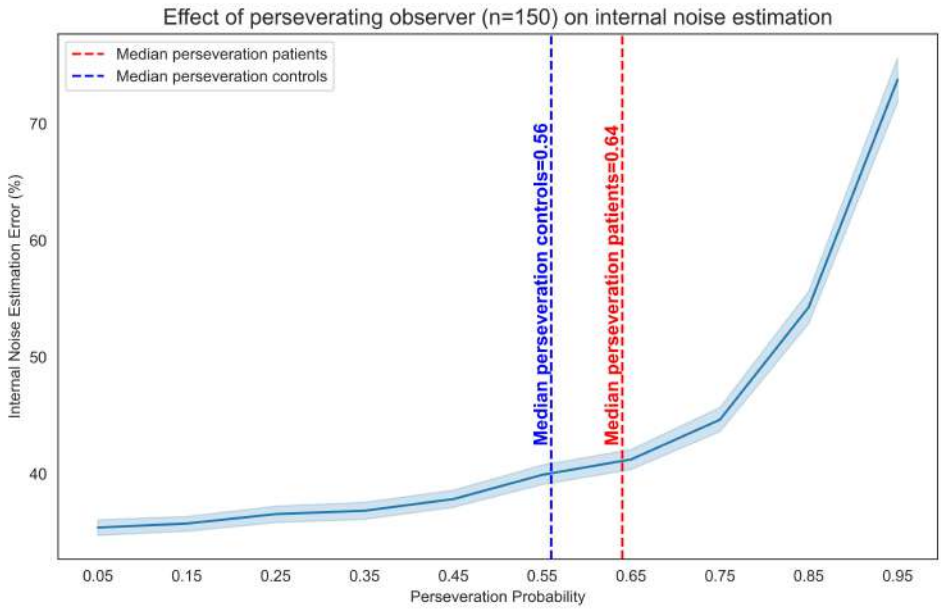


Fig. 5.7 **Effect of perseverative responding on internal noise estimation ($n = 150$) by Double-pass** Internal noise estimation error (%) as a function of perseveration probability in simulated observers with 150 trials. Estimation error remains relatively stable at lower levels of perseveration but increases sharply beyond a probability of 0.75. Vertical dashed lines represent the median perseveration probabilities for stroke patients (red, 0.64) and healthy controls (blue, 0.56), showing that patients are more likely to fall within a range where estimation becomes less reliable.

plify this bias, causing greater deviations between true and estimated noise. Additionally, higher β increases variability in estimates, as shown by the widening shaded regions in the simulations.

In sum, perseverative behaviour, which may be as common as $p \simeq 0.5$ in patients, has an impact on both kernel estimation (loosing up to 0.2 points of correlation to the true kernel) and internal noise (adding up to $10-20\%$ relative errors in extreme cases). While it is possible to identify and e.g., remove streaks of successive identical responses before computing such estimates, this method is likely to produce overestimates of perseveration and does not take the relation between kernel and stimuli into account when deciding if a trial is perseverated or not.

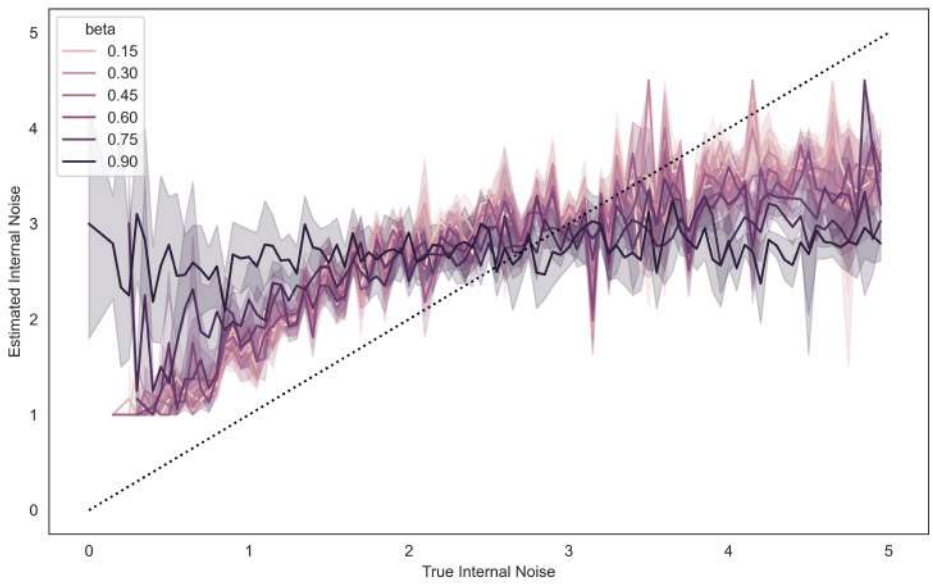


Fig. 5.8 **Impact of perseveration probability on internal noise estimation accuracy with Double-pass** Estimated versus true internal noise for simulated observers under varying levels of perseveration probability (β). Each curve (purple) represents a different β value, indicating the proportion of trials with repeated (perseverative) responses. At high β levels (e.g., $\beta \geq 0.75$), internal noise tends to be overestimated when the true internal noise is low, and underestimated when the true internal noise is high. This pattern reflects a systematic distortion in estimation induced by perseverative behavior. The dotted diagonal line indicates perfect estimation.

5.2 Problem statement

In this chapter, we have identified two key challenges that affect the reliability of our methods for estimating internal noise and mental representations: the limited number of trials and perseveration. The limited trial count imposes constraints on the precision of estimations, particularly that of internal noise. Perseveration disrupts both the stimulus-response relation and response variability, leading to large estimation errors in both kernel and noise. These two issues highlight critical limitations in our current approach, which compromise the interpretability of the preliminary results described in Chapter 4 and, more generally, the application of reverse correlation to patient populations.

Taken together, the results of Chapter 4 and 5 (Part II) lead to the following problem statement for the remained of this thesis:

In order to apply reverse correlation to patient populations and to fatiguable/perseverating stroke patients in particular, one needs to develop kernel and internal noise estimation methods that are both robust to low number of trials, and to local disruptions of decision strategy such as perseveration.

The next part of this thesis will introduce several methodological contributions that address this problem. In Chapter 6, we introduce and evaluate three new methods to estimate internal noise in the absence of double-pass trials. These methods address both problems above: first, because they do not rely on double-pass data, they allow using the complete set of trials in an experiment to infer internal noise, which is likely an advantage with small-trial setups such as here. Second, because they do not restrict internal noise measurements to specific blocks, they are also expected to be less sensitive to local perturbations such as perseverations. Of these three methods, two were developed by collaborators (Ladislas Nalborczyk and thesis director JJ Aucouturier) and one (GLM confidence intervals) by myself; my contribution is also to compare them against one another and against the double-pass method.

In Chapter 7, we introduce a new method to conjointly estimate both linear-observer parameters and perseverating episodes, using a joint model with two latent states (input-output hidden Markov model, or GLM-HMM). We show that this model is able to recover perseverating episodes by taking into account not only repeated responses but also stimuli-response relations and to improve the accuracy of kernel and noise estimates across non-perseverated episodes. In addition, the GLM-HMM method is compatible with the internal noise methods developed in Chapter 6, while not with the double-pass method (as perseverating episodes do not necessarily coincide with double-pass blocks).

In part IV, we then apply this new and improved set of methods to reanalyse the experimental data of Chapter 4 and confirm or infirm the clinical interpretability of reverse correlation estimates (Chapter 8). In addition, the accurate estimation of perseverating episodes allows a supplementary analysis of what factors may influence perseveration in patients, which we detail in Chapter 9.

Part III

Methodological contributions: new algorithms for the estimation of internal noise and perseveration

CHAPTER 6

Three new methods to estimate internal noise in the absence of double-pass measurements

The results of Chapter 4 and 5 (Part II) lead to identifying the need for internal noise estimation methods that are both robust to low number of trials and to perseverating patients, both aspects of our experimental conditions for which we showed that the classical double-pass procedure is not well-suited.

In the present chapter, we introduce and evaluate three new methods to estimate internal noise in the absence of double-pass trials. These methods address both problems above: first, because they do not rely on double-pass data, they allow using the complete set of trials in an experiment. Second, because they do not restrict internal noise measurements to specific blocks, they are also expected to be less sensitive to local perturbations such as perseverations.

Of these three methods, two were developed by collaborators (Ladislas Nalborczyk and thesis director JJ Aucouturier) and one (GLM confidence intervals) by myself; my

contribution also includes comparing them against one another and against the double-pass method.

This study was prepared in the format of a preprint, co-written with L.N, JJ.A. and M.V, and which is intended for submission at a methodological journal such as *Behaviour Research Methods* or *Quantitative Methods for Psychology*.

> Adl Zarrabi, A., Aucouturier, JJ., Nalborczyk, L. & Villain, M. (2025). *Three new analysis methods to estimate internal noise in data-driven experiments, in the absence of double-pass measurements*. Unpublished preprint.

We present here a verbatim of the manuscript, preceded by a short summary of the methods and main results.

6.1 Materials and methods

6.1.1 Baseline method: the double-pass procedure

The baseline method to evaluate internal noise from data-driven experiments is the so-called double-pass procedure, which presents a sequence of stimuli twice and measures the probability of agreement (p-agree) across repetitions (Burgess & Colborne, 1988; Lu & Dosher, 2008). Probability of agreement can then be converted into an equivalent level of internal noise σ_n using a backward model, which can be computed from computer simulations (see Chapter 3).

6.1.2 Alternative 1: the Intercept Method

In order to evaluate internal noise from data-driven data that do not include double-pass measurements, one first intuition is that, even if the data does not include exactly repeated trials, it does include a distribution of non-identical trials that vary in how much they resemble each other –some of them are quasi double-pass, others are more distant. By ranking all pairs of trials by increasing internal distance in stimulus space and measuring the consistency of responses over these pairs, we obtain a curve which, if

interpolated to pairs of trials with 0 internal distance, intersects the y-axis at a value that corresponds to the value of consistency that should have been observed for true double-pass trials, had such trials been observed (see manuscript Figure 2, below). The procedure involves a number of design parameters (e.g., whether the intercept is extrapolated with a polynomial or a logistic fit), which we describe and evaluate in the manuscript below.

6.1.3 Alternative 2: Accuracy method

As an alternative heuristic to estimate p_{agree} in the absence of double-pass data, we propose to estimate the participant's probability of agreement with an ideal, zero-noise, zero-bias observer responding to the same data with the same kernel as the participant. The rationale of this method is that, while double-pass experiments allow cancelling out the variability due to random trials (i.e., external noise) by repeating them identically, it is also possible to control for trial variability by simulating what a zero-noise observer with the same kernel would respond - in both cases, all remaining source of inconsistency should be attributed solely to internal noise. As above, the procedure involves a number of design parameters (e.g., whether accuracy is computed on all trials or only hits or correct rejections), which we describe and evaluate in the manuscript below.

6.1.4 Alternative 3: GLM Method

Finally, we already noted that an equivalent view of the linear-observer Eq. 3.9 is to consider responses r_t as the binary outcome of a generalized linear model (GLM), given by:

$$y^t = g(\beta_0 + \sum_{i=1}^{N} \beta_i x_i^t) \tag{6.1}$$

where the weights $\beta_{i=1...N}$ of the linear predictor correspond to the coordinates of kernel k, and input x_i correspond to stimulus data.

While a wealth of data-driven studies have used GLMs as a way to estimate kernels (Knoblauch & Maloney, 2008), the width of the confidence interval surrounding the β_is can also be interpreted as an indicator of internal noise. We therefore propose to estimate internal noise by estimating the 95% confidence intervals CI_i around GLM weights β_i

using their fitted standard errors backtransformed using the inverse of the link function g. Because GLM kernel values are expressed in units of internal-noise standard deviation (Murray, 2011), we normalize CI_i by the absolute value of the corresponding β_i, aggregate them across kernel dimensions in a single measure of confidence, and convert them back to standard-deviation estimates by multiplying by the square-root of the number of trials n. The procedure involves a number of design parameters (e.g., should we average, max or min the CI_i over all kernel dimensions?), which we describe and evaluate in the manuscript below.

6.1.5 PALIN Simulation methodology

To evaluate the methods, we used the PALIN simulation toolbox to implement the classical double-pass procedure to estimate observer internal noise, as well as the three Intercept, Accuracy and GLM methods described above. To evaluate each method, we simulated a range of linear observers with known, true internal values, let them encounter simulated experiments with varying numbers of trials, and compare each method's capacity to recover the true parameters. The process has inherent stochasticity, both in the observers (generated with random kernels), experiments (generated with random trials), and decisions (generated with random internal noise realization across trials). To account for this, we applied each method over a number of independent runs, and report average measures of accuracy and, when relevant, confidence intervals over runs and parameter settings.

6.2 Results

We compared the optimal variant of each method and their configurations (Intercept: GLM, no binning; Accuracy: all trials, weighted and non-weighted; GLM: 2% jitter, argmax/min, probit; see manuscript below for details) against each other and against the double-pass method, over their capacity to estimate true internal noise for simulated experiments.

At n=1000 trials, both the Intercept (Relative error:23%) and Accuracy (RE:25%)

methods provided better estimates than the traditional double-pass method (RE:33%). The GLM method (RE:38%) was less accurate as double-pass, but mostly because of its overestimation of low internal noise values. When restricted to $\sigma_n > 1$, GLM was more accurate than double-pass (RE:19%), and in fact both other methods.

Perhaps most importantly for our context here, at a low/very-low number of trials (ex. $n = 100$ single-pass or $n = 50$ repeated trials), the double-pass method evaluates with an unimpressive 48% relative error ([0.8–1.8] around the typical healthy participant estimate of 1.3 - Neri (2010)), while the best method (Accuracy) achieves an error of 30% (Manuscript Figure 10, reproduced below). Error rates below 20% are reached as early as 600 single-pass trials, using the GLM method under the assumption that $\sigma_n > 1$.

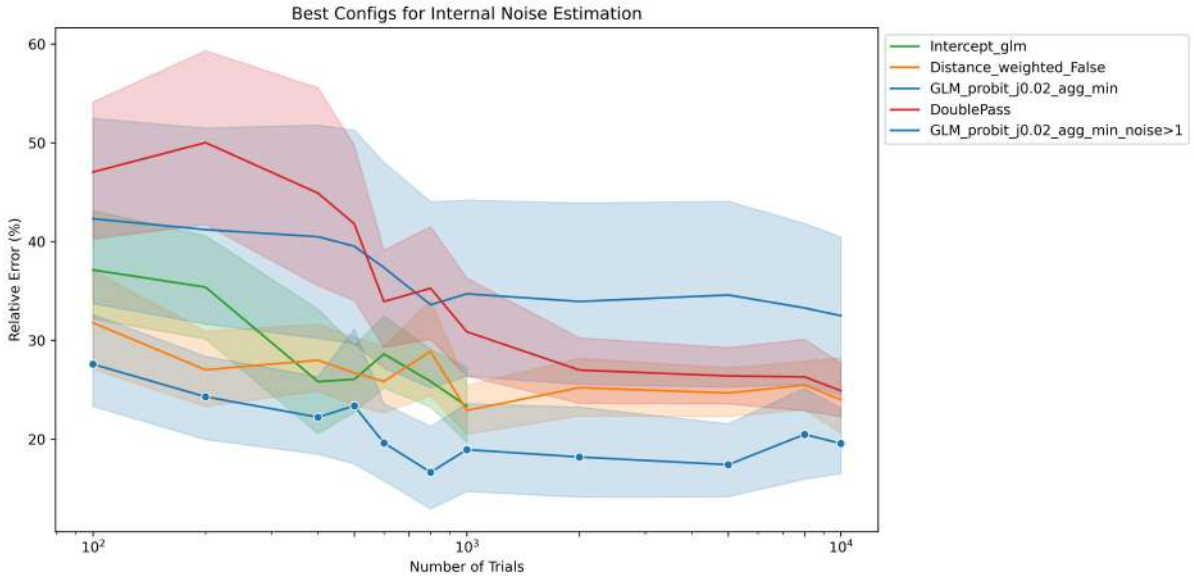


Fig. 6.1 i
n log scale]**Relative error of internal noise estimation across methods and trial counts [100, 10000] in log scale**. Intercept (`Intercept_glm`) and Accuracy (`Distance_not_weighted`) show consistently lower error than `Double_pass` across all trials counted up to 1000. While `GLM_probit_aggmin` overestimates internal noise at low levels, but it performs best among all methods when the noise level is restricted to $\sigma_n > 1$.

6.3 Study

Three new analysis methods to estimate internal noise in data-driven experiments, in the absence of double-pass measurements

Aynaz Adl Zarrabi[1], Jean-Julien Aucouturier[1*†], Ladislas Nalborczyk[2*†], Marie Villain[3*†]

[1*]Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, FEMTO-ST Institute, F-25000, Besançon, France.
[2]Aix Marseille Univ, CNRS, LPL.
[3]Paris Brain Institute (ICM), Inserm, CNRS, PICNIC-Lab, Sorbonne université, 75013, Paris, France.

*Corresponding author(s). E-mail(s): aucouturier@gmail.com; ladislas.nalborczyk@cnrs.fr; marie.villain@aphp.com;
Contributing authors: aynazadl1@gmail.com;
†Joint senior authorship.

Abstract

One of the methodological advantages of data-driven reverse-correlation experiments is to allow estimating not only the sensory/cognitive representations of participants, but also their *internal noise*. The dominant strategy to estimate internal noise from data-driven experiments is the double-pass procedure, in which responses to identically repeated trials are used to derive a measure of consistency. However, the double pass procedure is plagued with important practical limitations, among which its lack of precision for short experiments, its vulnerability to local perturbations, and its assuming that noise is stationary throughout the experiment. In this work, we introduce and evaluate 3 alternative analytical techniques to estimate internal noise from data-driven experiments, which do not require the availability of double-pass measurements and are therefore well-suited to short or continuous experimental designs. Using computer simulations, we show that the three techniques consistently outperform double-pass estimation for experiments with low number of trials (<1000), often more than halving percentage error. We make the three techniques available in an open-source Python/R toolbox and encourage the community to experiment with them.

Keywords: Internal noise, Reverse correlation, Double Pass

1 Introduction

While the idea of presenting an unknown system with random input sequences in order to estimate its transfer function can be traced back to Wiener-Volterra theories of system identification (Lee & Schetzen, 1965), its application to the modeling of cognitive systems has long been restricted to neurophysiological research, notably to characterize receptive fields of individual retinal (Marmarelis & Naka, 1972) or auditory neurons (Eggermont, Johannesma, & Aertsen, 1983). Starting possibly with the work of Ahumada and Lovell (1971), the technique was then adapted to human psychophysics, using randomized visual or auditory stimulus in input, and taking behavioral choices (e.g., yes/no responses) instead of neuronal spikes as the systems' output variables. In the process, the procedure has taken on a number of alternative names, incl. *reverse correlation* (Ringach & Shapley, 2004), *classification images* (Murray, 2011), *bubbles* (Gosselin & Schyns, 2001), *response-triggered averaging* (Owen Brimijoin, Akeroyd, Tilbury, & Porr, 2013) or, perhaps most inclusively, *data-driven methods* (Adolphs, Nummenmaa, Todorov, & Haxby, 2016). In the auditory domain, data-driven experiments were used to study low-level sensory processes such as the detection of tones in noise (Ahumada & Lovell, 1971) or loudness weighting (Ponsot, Susini, Saint Pierre, & Meunier, 2013), language processes such as speech intelligibility (Venezia, Hickok, & Richards, 2016) or phoneme categorization (Varnet, Wang, Peter, Meunier, & Hoen, 2015), as well as higher-level decision tasks such as social prosody (Ponsot, Burred, Belin, & Aucouturier, 2018) or the vocal detection of sleepiness (Thoret, Andrillon, Gauriau, Leger, & Pressnitzer, 2024). In the visual domain, data-driven methods have also found wide application not only for low-level detection and discrimination tasks such as line detection (Neri & Heeger, 2002) or letter identification (Solomon & Pelli, 1994), but also for face recognition (Mangini & Biederman, 2004), emotional expressions (Jack, Garrod, Yu, Caldara, & Schyns, 2012) or social traits (Dotsch & Todorov, 2012).

A prominent model to understand data-driven methods is the *linear observer* model (Lu & Dosher, 2008; Murray, 2011). Consider a 2-alternative forced-choice (2AFC) experiment where two randomized signals s_1^t and s_2^t are presented in each trial t, and an observer is tasked to identify which of s_1,s_2 best matches an internal template k (typically called a *kernel*, after Wiener's theory). A minimal model for how such an observer may come to a decision is to compute decision variables d_1^t and d_2^t by taking the dot product of the stimuli with the kernel $s \cdot k$, and add realizations from an independent source of 'internal noise' n (called 'internal' in contrast to 'external noise', which is the noise applied experimentally to the stimuli s_1 and s_2):

$$d_1^t = s_1^t \cdot k + n_1^t \tag{1}$$

$$d_2^t = s_2^t \cdot k + n_2^t \tag{2}$$

The model assumes that the observer identifies the signal as s_2 if d_2 plus some constant b (response bias) is larger than d_1:

$$(s_2^t - s_1^t) \cdot k + (n_2^t - n_1^t) > b \tag{3}$$

or, equivalently,

$$r^t = \begin{cases} 2 & \text{if } (s_2^t - s_1^t) \cdot k + n^t > b \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

where n is a source of internal noise with a variance σ_n that is twice the variance of n_1 and n_2 (for a review of alternative formulations of observer models, see Lu and Dosher (2008)). Based on this model, data-driven methods provide procedures to estimate both kernel k, internal noise σ_n and bias b from a series of stimulus-response pairs $(s_1^t, s_2^t; r^t)$ corresponding to how a given observer responded to a given experiment.

Importantly, while a majority of studies focus on analysing observer kernels k, often describing them as the primary (linear) mechanism behind the observed process (Adolphs et al., 2005), the associated internal noise σ_n isn't just a waste-basket parameter capturing all remaining nonlinearities in the process (i.e. the residual to an imperfect linear fit), but rather a genuine property of the biological system. It is indeed well-established that stochasticity is pervasive at all levels of biological information-processing systems, from thermodynamic/quantum noise acting on sensory receptors (e.g. photons hitting retinal photoreceptors at random rates), variability in the strength and timing of neuronal action-potential transmission (e.g. 'background' spontaneous neurotransmitter release in synapses), up to macroscopic variability in response or motor behaviour (Faisal, Selen, & Wolpert, 2008). In other words, the repetition of identical stimulation to a participant does *not* generate identical responses, and that amount of variability is a mechanistic property of the sensory-cognitive process in the same way that the kernel is (Neri, 2010). For instance, phenomena such as stochastic resonance show that the addition of an optimal level of noise to a weak input can enhance its detectability (Russell, Wilkens, & Moss, 1999), and different internal noise levels may be the only thing that separates sensory processes with otherwise similar kernels or sensitivity (e.g. luminance and contrast perception - Allard and Faubert (2006)). Previous studies that estimate internal noise from data-driven paradigms have for instance documented that it is statistically independent from dprime sensitivity Neri (2010); that its population-average level is similar across sensory modalities (Neri, 2010); that it is consistent across tasks intra-individually Vilidaite and Baker (2017); that it increases with aging (Yan et al., 2020) and that it can be modulated with repeated transcranial magnetic stimulation (rTMS) in sensory cortices (Vilidaite & Baker, 2018). At the clinical level, abnormalities in internal noise level have been associated with deficits of language prosody perception after a brain stroke (Adl Zarrabi et al., 2024), or with autism spectrum disorders (Park, Schauder, Zhang, Bennetto, & Tadin, 2017; Vilidaite, Yu, & Baker, 2017). In many experiments, it is therefore at least as important to estimate internal noise as the kernel.

The dominant analytical strategy to evaluate internal noise from data-driven experiments is the so-called *double-pass* procedure. Because response variability to identical stimuli in the linear observer (Eq.4) can only be attributed to internal noise, the

double-pass procedure presents a sequence of noisy stimuli multiple times (typically, twice) and measures the consistency of responses as a probability of agreement across repetitions (Burgess & Colborne, 1988; Lu & Dosher, 2008). Probability of agreement can then be converted into an equivalent level of internal noise σ_n using computer simulations of the linear model (see e.g. Goupil, Ponsot, Richardson, Reyes, and Aucouturier (2021) and *Methods*, below). While double-pass consistency is considered the easiest and most direct way to estimate internal noise in data-driven studies, with regular methodological discussion on how to best implement it (Hasan, Joosten, & Neri, 2012) or what precision it can achieve (Adl Zarrabi et al. (2024)-Supplemental info), the procedure remains plagued with important practical limitations. First, the precision of double-pass consistency increases with the number of repeated stimuli, and duplicating stimuli mechanically adds duration to the experiment. This may be impractical when working with fatigable patients (Adl Zarrabi et al., 2024) or in large-N experimental designs with online participants. Second, double-pass consistency is fragile to local perturbations in the experiment. For instance, technical problems or participant distraction occurring during a double-pass block can be eliminated from kernel estimates by discarding those trials, but will perturbate the complete measure of internal noise. Third, the procedure can only provide a global measure of internal noise over the experiment (ex. comparing the first and last block of trials) and does not allow measuring continuous changes of internal noise in situations where it cannot be assumed to be stationary, or in response to experimental manipulations. Alternative procedures have been introduced to estimate internal noise that do not require double-pass measurements (e.g. equivalent noise method - Pelli (1985); pedestal masking - Vilidaite and Baker (2017)), but they require specific experimental designs (typically, staircase procedures) and suffer from the same limitations.

In this work, we introduce and evaluate 3 alternative analytical methods to estimate internal noise from data-driven experiments, which do not require the availability of double-pass measurements or specific staircase procedures, and are therefore well-suited to short and/or continuous experimental designs. The first two strategies are based on consistency metrics that approximate double-pass probability of agreement using non-identically repeated trials, while the latter is a direct evaluation of standard deviation around the kernel. For each, we evaluate the impact of key design parameters, and offer a comparison of their performance to that of the double-pass procedure using simulation data. We also provide Python and R-language implementations of the new methods in the form of a novel open-source toolbox[1].

2 Analysis techniques for the estimation of internal noise

2.1 Baseline method: Double-pass consistency and simulated backward models

The baseline method to evaluate internal noise from data-driven experiments is the so-called *double-pass* procedure, which presents a sequence of stimuli twice and measures

[1] https://github.com/neuro-team-femto/palin

the probability of agreement p_{agree} across repetitions (Burgess & Colborne, 1988; Lu & Dosher, 2008). Probability of agreement can then be converted into an equivalent level of internal noise σ_n using a backward model, which can be computed from computer simulations.

In more details, the ideal observer model of Eq. 4 provides a generative model by which one can simulate the responses given by an observer with a given kernel and internal noise level. From these responses to simulated experiments, one can evaluate an empirical probability of agreement p_{agree} over n repeated trials. This provides a *forward model* that associates an observed p_{agree} to the (true, known) value of an observer's internal noise value σ_n under the linear observer assumption. This model is stochastic: for a given observer, and a given experiment, responses have stochasticity because of internal noise realizations, and p_{agree} therefore lies on a probability distribution for each value of σ_n.

The task of estimating internal noise from a real observer's response is the corresponding *backward model*: from an empirical observation of p_{agree} over n repeated trials, find the most plausible internal noise level $\hat{sigma}_n$ that would forward-generate this probability. To extend to the case where observers are biased (i.e a non null parameter b in Eq. 4), one can also observe probability of answering the first interval (in a 2AFC design) p_{first}. In this generalized situation, the forward model maps two hidden variables (internal noise σ_n and observer bias b) to two observations (p_{agree} and p_{first}).

A typical practical implementation of this procedure is to simulate observers with an arbitrary kernel (ex. a simple scalar) and a grid of (true, known) internal noise σ_n and bias b values and let them encounter a single large (e.g. $n = 1000$ trials or more) experiment for several runs (e.g. 100 or 1000); computed the average value of p_{agree} and p_{first} over several realizations of the same simulated experiment for each value of σ_n and b; and store this in a lookup table. Given the empirical observation of probabilities p_{agree} and p_{first} for a real observer, the procedure can then find the pair of internal noise $\hat{sigma}_n$ and bias $\hat{b}$ that correspond to the closest pair of probabilities in the table.

Figure 1 illustrates typical lookup data computed using a simulated linear observers with $\sigma_n \in [0, 5]$ and $b \in [-5, 5]$, expressed in units of external (stimulus) noise. Empirical probabilities p_{agree} and p_{first} are computed over simulated double-pass experiments with $n = 10^4$ repeated trials, and averaged over 10 realizations. For unbiased observers, σ_n maps non-linearly but bijectively to p_{agree}. However, symmetry of p_{agree} for positive and negative biases b illustrates the need to include p_{first} to disambiguate the underlying values of b and σ_n. Although authors have routinely used simulated forward models to estimate internal noise from double-pass data-driven experiments (Goupil et al., 2021; Neri, 2010; Ponsot et al., 2018), it is to be noted that p_{agree} and p_{first} also have analytical asymptotic formulations in the $n \to \infty$ case, which can also be used to compute a backward lookup table (see Appendix A).

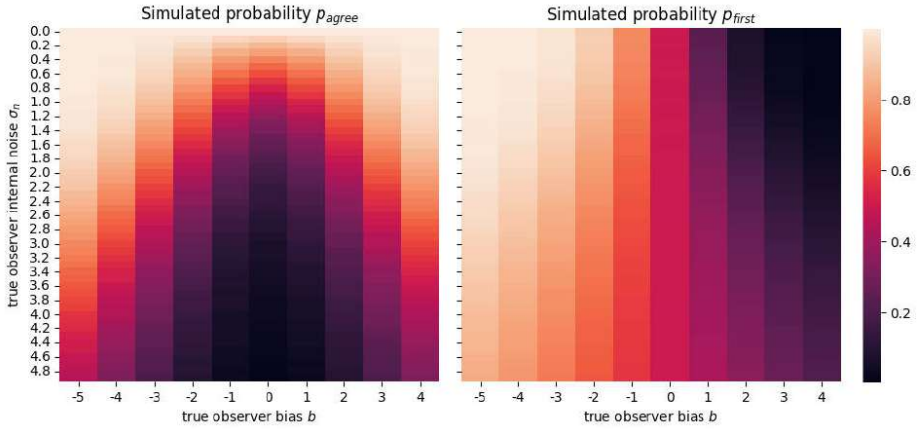


Fig. 1 Lookup data computed using simulated linear observers which show the influence of internal noise $\sigma_n \in [0, 5]$ and decision criterion $b \in [-5, 5]$ on p_{agree} and p_{first} S(forward model) Left: The simulated p_{agree} is highest when internal noise is low and bias is large (either positive or negative) and it is lowest when the bias is zero and internal noise is high. The plot is symmetric around $b = 0$, which illustrates the need of using p_{first} to disambiguate the direction of bias. Right: The simulated p_{first} increases monotonically with bias and captures directional trends in responses independently of stimuli. Lower internal noise and more positive bias associated with a higher p_{first}.

2.2 Alternative 1: Intercept method

In order to evaluate internal noise from data-driven data that do not include double-pass measurements, one first intuition is that, even if data does not include exactly repeated trials, it does include a distribution of non-identical trials that vary in how much they resemble each another - some of them quasi double-pass, other more distant. By ranking all pairs of trials by increasing internal distance in stimulus space, and measuring the consistency of responses over these pairs, we obtain a curve which, if interpolated to pairs of trials with 0 internal distance, intersects the y-axis at a value that corresponds to the value of consistency that should have been observed for true double-pass trials, had such trials been observed.

In more details, using response data from a n-trial 2AFC experiment, we consider all $\frac{n(n-1)}{2}$ combinations of trials i and j, compute their vector difference $\vec{d_{i,j}}$ as

$$\vec{\Delta_{i,j}} = ((s_1^i - s_2^i) - (s_1^j - s_2^j)) \tag{5}$$

and use it to derive a measure of internal distance, either as the L2-norm (i.e., how similar the two trials are physically) or, if available, by dot-product projecting it on an estimate of the participant's kernel k (i.e. how similar the trials' decision variables

are for this participant):

$$d(i,j) = \begin{cases} \texttt{RMS}(\vec{\Delta_{i,j}}) \\ |\Delta_{i,j} \cdot \ k| \end{cases} \tag{6}$$

where RMS is the root-mean-square of the vector coordinates of the trial difference.

For each pair of trials (i, j) with internal distance $d(i, j)$, we then compute response agreement $r_{i,j}$ as

$$r_{i,j} = \begin{cases} 1 & \text{if } r^i = r^j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where r^i is the trial response defined as Eq. 4. Finally, we estimate p_{agree} as the y-axis intercept of a parametric curve fitted on the graph of $r_{i,j}$ against $d(i, j)$. To do so, we either group trial pairs of similar internal distance $d(i, j)$ using $k = 1 \dots n_{bins}$ (n_{bins} 50-100, with an optimal value to be determined) and compute for each bin k the probability of agreement p^k_{agree} as the average agreement among all pairs in k

$$p^k_{agree} = \frac{1}{|\Omega_k|} \sum_{(i,j)\in\Omega_k} r_{i,j} \tag{8}$$

where Ω_k is the set of all trial pairs (i, j) in bin k, or do not bin and consider all $\frac{n(n-1)}{2}$ pairs and fit the raw binary agreements $r_{i,j}$. In practice, the distribution of $r_{i,j}$ and p^k_{agree} along increasing $d(i, j)$ is roughly sigmoid-shaped (Fig. 2) and to estimate the intercept, we propose to either fit an order-3 polynomial $p(d(i, j))$ and compute $p(0)$, or to fit a generalized linear model (GLM) with a logistic link function and extract the intercept of the linear model (in probit space).

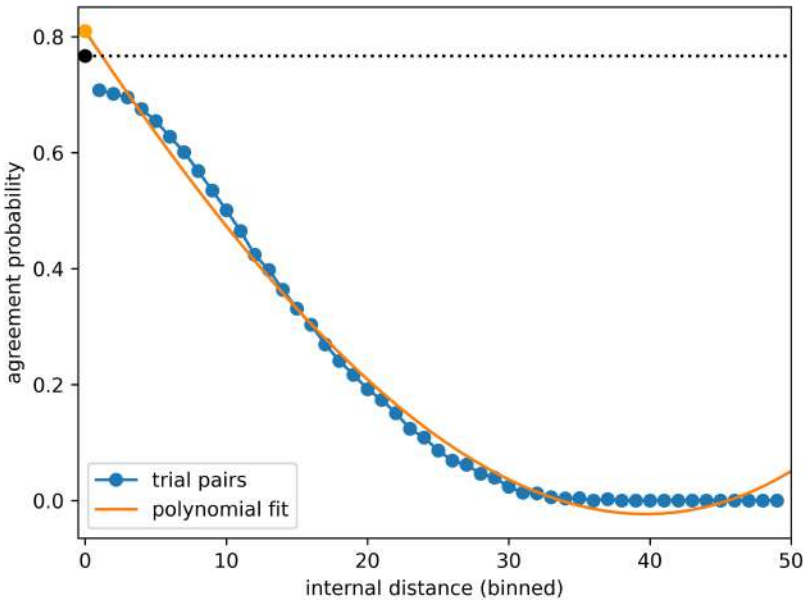


Fig. 2 Estimating probability of agreement at different bin levels for trial pairs using polynomial interpolation. Trial pairs, generated a simulated 1000-trial experiment with an observer at $\sigma_n = 1$, are grouped by increasing internal distance (x-axis) and agreement probabilities are computed per bin (y-axis). A degree-3 polynomial is fitted to the binned values (orange), and its y-axis intercept provides an estimate $\hat{p}_{\text{agree}}$. Black dot marks true p_{agree} (0.76) observed in the simulated experiment (1000 double-pass trials)

Figure 2 illustrates the typical application of the procedure on illustrative simulated data, where the observer's true internal noise value is known and controlled (here, $\sigma_n = 1$). Trial pair data $d(i, j)$ are binned with $n_{bins} = 50$, and agreement between pairs is averaged as probabilities of agreement in each bin. The y-intercept of a polynomial fit to these data gives an estimate of $\hat{p}_{agree}$ for the unobserved case that $d(i, j) = 0$, which can be compared with p_{agree} estimated over double-pass trials of the same experiment (here, an empirical value $p_{agree} = 0.76$ measured on 1000 repeated trials). As in the double-pass case, p_{agree} can then be converted to internal noise value σ_n (and bias b) with a backward lookup model (see Section 2.1).

In sum, the Intercept method provides a heuristic to estimate p_{agree} over unobserved repeated trials, by extrapolating the agreement of pairs of imperfectly repeated trials over the actual experiment. The procedure involves a number of design parameters, namely whether trial pair distance $d(i, j)$ is computed as physical trial distance or as a projection on the participant's kernel, whether agreement data is binned as

p^k_{agree} or left as r_i, j, and whether the intercept is computed with a polynomial or GLM fit. In the following, we evaluate these design choices against simulated data, and provide a recommendation for the parameters that provide the best estimates.

2.3 Alternative 2: Accuracy method

As an alternative heuristic to estimate p_{agree} in the absence of double-pass data, we propose to estimate the participant's probability of agreement with an ideal, zero-noise, zero-bias observer responding to the same data with the same kernel as the participant. The rationale of this method is that, while double-pass experiments allow to cancel out the variability due to random trials (i.e., external noise) by repeating them identically, it is also possible to control for trial variability by simulating what a zero-noise observer with the same kernel would respond - in both cases, all remaining source of inconsistency should be attributed solely to internal noise.

In more details, for a given set of n single-pass trials (s^t_1, s^t_2) and participant responses r^t, we compute the participant's kernel k - typically, using the classification-image method of computing the average of stimuli chosen as one response option, subtracted with the average of stimuli chosen as the other option (Murray, 2011). We then compute a binary variable $a(t)$ that describes the agreement of participant responses r^t with the zero-noise, zero-bias decision criteria $(s^t_2 - s^t_1) \cdot k$

$$a(t) = \begin{cases} 1 & \text{if } (s^t_2 - s^t_1) \cdot k > 0 \text{ and } r^t = 2, \text{ or} \\ & \text{if } (s^t_2 - s^t_1) \cdot k < 0 \text{ and } r^t = 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We then compute the probability of agreement $P[a(t) = 1]$ over n (or, equivalently, the mean accuracy of response $\frac{1}{n}\sum_t\ a(t)$), and take this measure as an estimate of $\hat{p}_{agree}$:

$$\hat{p}_{agree} = P[a(t) = 1] = \frac{1}{n}\sum_t\ a(t) \tag{10}$$

.

As an alternative, we investigate two variants of the accuracy procedure: first, instead of computing $P[a(t) = 1]$ for all trials, we consider computing it only for hits (conventionally defined as $r^t = 2$ when $(s^t_2 - s^t_1) \cdot k > 0$), i.e. $P[a(t) = 1/(s^t_2 - s^t_1) \cdot k > 0]$), or only for correct rejections $P[a(t) = 1/(s^t_2 - s^t_1) \cdot k < 0]$). Second, instead of simply averaging $a(t)$ over all trials, we consider weighting it by the trial's (absolute) activation magnitude, i.e.

$$\hat{p}_{agree} = \frac{\sum_t\ |(s^t_2 - s^t_1) \cdot k|.a(t)}{\sum_t |(s^t_2 - s^t_1) \cdot k|} \tag{11}$$

with the rationale that agreement is more important on trials where the response if normally unambiguous.

In sum, the Accuracy method provides a heuristic to estimate p_{agree} over single-pass trials by computing the agreement of the participant's response with a zero-noise,

zero-bias observer over the same trials. The procedure involves a number of design parameters, namely whether accuracy is computed on all trials or only hits or correct rejections, and whether accuracy is weighted by trial activation. In the following, we evaluate these design choices against simulated data, and provide a recommendation for the parameters that provide the best estimates.

2.4 Alternative 3: GLM Method

An equivalent view of Eq. 4 is to consider responses r_t as the binary outcome of a logistic regression (or, equivalently, a generalized linear model - GLM), given by:

$$y^t = g(\beta_0 + \sum_{i=1}^{N} \beta_i x_i^t) \tag{12}$$

where the weights $\beta_{i=1...N}$ of the linear predictor correspond to the coordinates of kernel k, and input x_i correspond to stimulus data (in the 2AFC case of Eq. 4, x_i^t is the i^{th} coordinate of stimulus difference $s_1^t - s_2^t$) - both of dimension N; and g a non-linear link function (logit or probit). While a wealth of data-driven studies have used GLMs as a way to estimate kernels (Knoblauch & Maloney, 2008), the width of the confidence interval surrounding the β_is can also be interpreted as an indicator of internal noise.

In more details, we propose to estimate the internal noise of an experimental participant by fitting their experimental data with a probit/logit taking as input the N-dimensional vector of stimulus difference $s_1^t - s_2^t$ (in the case of 2AFC), and as output the binary outcome 0/1 of which stimulus was chosen. We then estimate the 95% confidence intervals $\mathcal{CI}_i$ around GLM weights β_i using their fitted standard errors backtransformed using the inverse of the link function (a procedure implemented as `stats.confint(model)` in R or `model.conf_int()` in Python `statsmodel`). Because GLM kernel values are expressed in units of internal-noise standard deviation (Murray, 2011), we then normalized confidence interval width $\mathcal{CI}_i$ by the absolute value of the corresponding β_i

$$\hat{\mathcal{CI}}_i = \frac{\mathcal{CI}_i}{|\beta_i|} \tag{13}$$

This procedure results in N estimates of variability, around each dimension of the kernel. Because internal noise is typically modeled as a one-dimensional distribution acting on the (scalar, dot-product) decision variable rather than on individual kernel components (Eq. 4), we therefore aggregate the N estimates of $\hat{\mathcal{CI}}_i$ using an statistics, to be determined among:

- `min`: $\hat{\mathcal{CI}} = \min_i \hat{\mathcal{CI}}_i$
- `max`: $\hat{\mathcal{CI}} = \max_i \hat{\mathcal{CI}}_i$
- `mean`: $\hat{\mathcal{CI}} = \text{mean}_i\ \hat{\mathcal{CI}}_i$
- `median`: $\hat{\mathcal{CI}} = \text{median}_i\ \hat{\mathcal{CI}}_i$
- `argmin`: $\hat{\mathcal{CI}} = \hat{\mathcal{CI}}_j$ where $j = \text{argmin}_i \beta_i$
- `argmaxn`: $\hat{\mathcal{CI}} = \hat{\mathcal{CI}}_j$ where $j = \text{argmax}_i \beta_i$
- `argmedian`: $\hat{\mathcal{CI}} = \hat{\mathcal{CI}}_j$ where $j = \text{argmedian}_i \beta_i$

Finally, we convert confidence intervals (essentially standard errors) to estimates of standard deviation, by multiplying them by the square-root of the sampling unit (number of trials n), taking this value as the estimate of observer internal noise $\hat{\sigma}_n$.

$$\hat{\sigma}_n = \hat{\mathcal{CI}}\sqrt{n} \tag{14}$$

Figure 3 illustrates the typical application of the procedure on illustrative simulated data, where the observer's true internal noise value is known and controlled, using the argmax aggregation strategy and for a range of number of trials n. The figure illustrates that $\hat{\mathcal{CI}}$ estimates typically scale linearly with true internal noise values (Figure 3-bottom left) and non-linearly as the square root of n, while $\hat{\sigma}_n$ corrects this dependency (Figure 3-bottom right). The figure also illustrates that the method tends to overestimate CIs for small values of internal noise in the $[0, 1]$ range (see also Figs. 7 and 8). This is typically the consequence of numerical errors for standard error around β_i estimates when training data involves too little variability (at the boundary case of $\sigma_n = 0$, Python GLM estimates will fail numerically with a co-called `PerfectionSeparation` error). While such situations are unlikely to occur with experimental data, as typical internal noise value in human observers across a variety of tasks is 1.3 (Neri, 2010), and even larger for patients (Adl Zarrabi et al., 2024), we investigate a mitigation strategy that adds small quantities of jitter to response data, i.e. randomly inverting a small percentage of binary responses before fitting the GLM.

Because the procedure provides a linear correlate of internal noise, but empirically does not have unitary slope (Figure 3-bottom left), a final step is to rescale $\hat{\sigma}_n$ using a linear model (ordinary least-square regression) $\sigma_n \sim \hat{\sigma}_n + \text{n\_trials}$ trained on simulated data for a range of observers with known internal noise values and experiments with various `n_trials`. The model can then be saved and reused on empirical estimates of $\hat{\sigma}_n$ (see Section *Methods* for details).

In sum, the method provides a direct model-based way to estimate internal noise on non double-pass data via the proxy of normalized, trial-scaled confidence intervals around GLM weights. The procedure involves a number of design parameters, namely the aggregation function over multiple $\hat{\mathcal{CI}}_i$, the amount of jitter, and the GLM link function (logit/probit). In the following, we evaluate these design choices against simulated data, and provide a recommendation for the parameters that provide the best estimates.

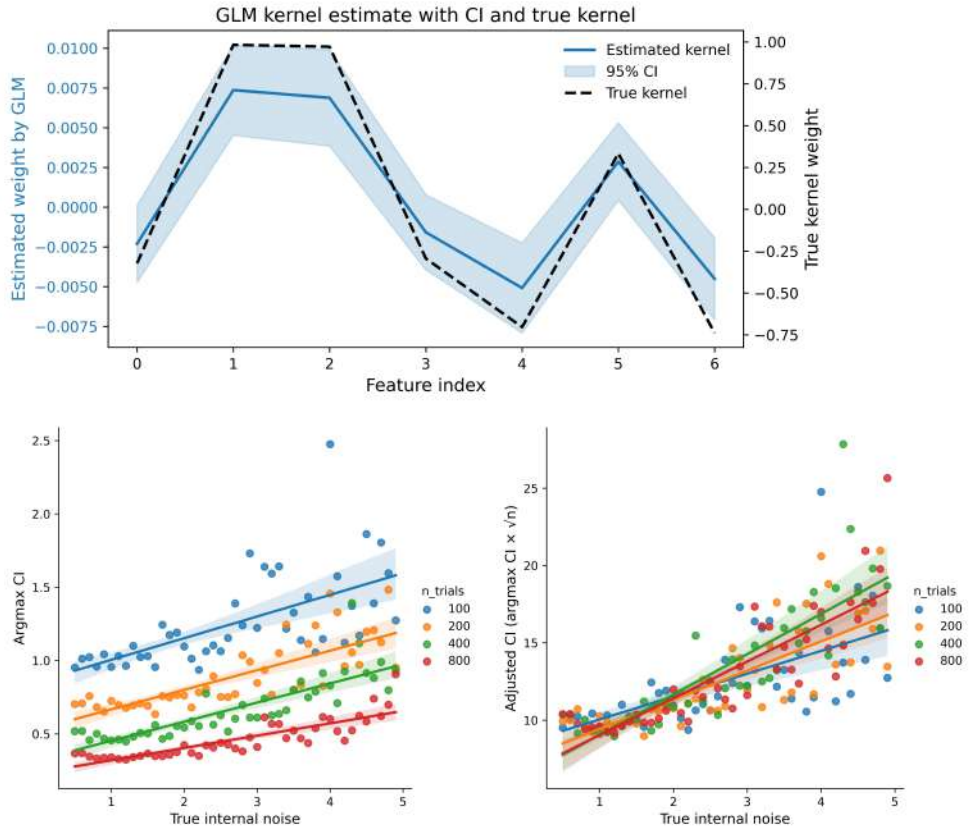


Fig. 3 GLM kernel estimate and normalized confidence intervals as a function of internal noise and number of trials, illustrated on simulated data. Top: The GLM (probit) can accurately recover the shape of the true kernel of an observer. The 95% confidence interval (CI) around each estimated feature weight is shown. CI width varies across features. To summarize uncertainty in a single value, we focus on the feature with the strongest weight and extract its CI width (referred to as the `argmax` aggregation method). **Bottom left:** Estimated values of $\hat{CI}$ increase with increasing internal noise, as expected. However, $\hat{CI}$ also depends non-linearly on the experiment's number of trials: more trials result in lower $\hat{CI}$s. **Bottom right:** After adjusting for the number of trials by multiplying the CI width by the square root of n, a linear relationship emerges between internal noise and adjusted CI.

3 Methods

3.1 Simulation toolbox

To evaluate the methods, we developed an open-source Python/R toolbox (`PALIN`) that allows simulating linear observers whose decision-making processes follow the

formulation described in Equation 4. The toolbox enables the flexible simulation of reverse-correlation experiments by varying key parameters such as internal noise σ_n, decision bias b, and the observer kernel k. It also allows control over the level of external noise, defined as the standard deviation of the Gaussian distribution from which each stimulus is drawn. Users can specify the number of stimulus dimensions and select from different experimental designs, including single-pass and double-pass. The toolbox also supports different stimulus presentation formats, such as 1AFC and 2AFC tasks. The toolbox is implemented in both the Python and R programming languages, and available open-source at https://github.com/neuro-team-femto/palin.

Using PALIN, we implemented the classical double-pass procedure to estimate observer internal noise, as well as the three Intercept, Accuracy and GLM methods described above. To evaluate each method, we simulate a range of linear observers with known, true internal values, let them encounter simulated experiments with varying numbers of trials, and compare each method's capacity to recover the true parameters. The process has inherent stochasticity, both in the observers (generated with random kernels), experiments (generated with random trials), and decisions (generated with random internal noise realization across trials). To account for this, we apply each method over a number of independent runs, and report average measures of accuracy and, when relevant, confidence intervals over runs and parameter settings. We describe below the parameters of each simulation.

3.2 Evaluation of Intercept design parameters

The implementation of the Intercept method (Section 2.2) involves several design parameters, which we evaluated against each other before choosing the best candidate for comparison with other methods.

First, we evaluated the impact of how many bins are used to discretize trial-pair distances, in both the poynomial and GLM fitting methods. To do so, we ran computer simulations of ideal observers encountering simulated double-pass experiments (n=1000 random trials, repeated n=1000 trials) for a range of observer internal noise values ($[0, 5]$). We then extracted both estimations of probability of agreement ($\hat{p}_{agree}$) over the 1000 repeated trials using the double-pass method (Section 2.1), and estimations of $\hat{p}_{agree}$ over the first non-repeated 1000 trials using the Intercept method. We compared 24 variants of the Intercept method (12 number of bins in $\{10, 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 5000, 10000\}$, 2 fitting methods: polynomial or GLM). For each parameter combination, we compared the evolution of estimated Intercept $\hat{p}_{agree}$ against internal noise against the estimated Double-pass $\hat{p}_{agree}$, taken as a baseline, using the root-mean-square error (RMSE). To alleviate stochasticity due to internal noise, each combination of parameters (observer internal noise, number of bin, fitting method) was simulated 10 times. RMSE was computed on the average $\hat{p}_{agree}$ across these 10 runs.

Second, we selected the number of bins that lead to the minimal RMSE value of $\hat{p}_{agree}$ against internal noise for both polynomial and GLM fitting methods, and compared the performance of this optimal number of bins to variants of the Intercept method that did not use any binning (i.e., fitted intercept to the complete number of trial pairs). To do so, we ran another 10 runs of 1000-trial double-pass simulations

for a range of observers with internal noise in $[0, 5]$, comparing double-pass estimations $\hat{p}_{agree}$ on 1000 repeated trials and Intercept estimations $\hat{p}_{agree}$ for 4 variants (binned/non-binned, polynomial/GLM). We evaluated Intercept variants on the basis of their RMSE to double-pass $\hat{p}_{agree}$. The variant with minimal RMSE was kept for later comparison against other methods (see Section 3.5 below).

3.3 Evaluation of Accuracy design parameters

The design parameters for the Accuracy method (section 2.3) include, first, which subsets of trials to use when computing accuracy (all trials, only hits, or only correct rejections) and, second, whether accuracy should be weighted by activation magnitude.

To evaluate these design parameters, we ran simulations of observers performing $n = 1000$ trials simple-pass experiment (no repeated trials) across a range of internal noise values($[0, 5]$). Observer kernels used for computing accuracy were estimated as if in real experimental data using the classical classification image method, which computed the average of stimuli chosen as one response option and subtracting the average of stimuli chosen as the other option (Murray, 2011). A total of six configurations were tested; three different trial masks (all trials, hits, CRs) crossed with two weighting schemes (weighted vs unweighted). Each simulation was repeated 10 times to average out variability. The Accuracy variant with the lowest average RMSE to double-pass $\hat{p}_{agree}$ was selected for subsequent comparisons with other methods (see Section 3.5 below).

3.4 Evaluation of GLM-method design parameters

The GLM method (section 2.4) involves several design variables, including the choice of aggregation function (`min`, `max`, `mean`, `median`, and several "arg" statistics (`argmax`, `argmedian`, `argmin`), the GLM link function (`logit` - log-odds, and `probit` -inverse-Gaussian CDF), as well as the level of jitter used to mitigate small-noise overestimation. Contrary to the preceding Intercept and Accuracy method, which estimates a proxy of $\hat{p}_{agree}$ which is then subjected to a lookup table, the GLM method provides a direct estimate $\hat{\sigma}_n$ of internal noise (corrected for slope with a pretrained linear regression, see Section 2.4).

We first evaluated the influence of jitter level on the method's evaluation of internal noise. To do so, we ran computer simulations of ideal observers encountering simulated single-pass experiments ($n = 1000$ random trials) for a range of observer internal noise values ($[0, 5]$). We then extracted estimations of internal noise with GLM methods computed over a range of jitter values $\{1\%, 2\%, 5\%, 10\%, 15\%, 20\%\}$ and with fixed aggregation mode (`argmax`) and link (`logit`). For each configuration, we then computed the averaged estimated internal noise value over 10 runs. We compared jitter variants on the basis of their RMSE to the first diagonal (corresponding to true internal noise values). The jitter value with minimal RMSE was kept for later comparison against other variants.

Second, we compared variants across aggregation mode and link function choices, while fixing jitter at its optimal value. Over a large range of simulation parameters, aggregation modes `max`, `mean` and `argmin` led to missing values in the estimation

of confidence intervals and were discarded from the comparison. For the remaining 8 configurations (`argmax`,`argmedian`,`median` and `min`; link `logit` and `probit`), we let ideal observers with internal noise values $[0, 5]$ encounter simulated single-pass experiments ($n = 1000$ random trials), and computed GLM internal noise estimates over 10 runs. As before, we compared variants on the basis of RMSE to the first diagonal. The GLM variant with the lowest average RMSE was selected for subsequent comparisons with other methods (see Section 3.5 below).

In both evaluations, we used linear regression models to correct for non-unitary slope in the relation between $\hat{\sigma}_n$ and true internal noise (see section 2.4). To ensure a fair comparison of all simulation settings, we trained a separate model for every combination of aggregation mode and link function, and used that model for all jitter values. Each model was estimated using simulations of observers with internal noise $\in [0, 10]$, single-pass experiments with a number of trials $\in [100, 2000]$ and fitted over 10 independent runs.

3.5 Comparison of internal noise estimations across methods

Finally, we compared the optimal variants of each method over their capacity to estimate true internal noise. To do so, we trained separate lookup models to convert $\hat{p}_{agree}$ estimates from each of the three consistency methods (double-pass, intercept, accuracy). The models were estimated using simulations of observers with true internal noise $\in [0, 10]$ and bias $\in [-5, 5]$, encountering single- (Intercept, Accuracy) or double-pass (Double-pass) experiments with $n = 1000$ trials, and averaged over 10 runs. GLM estimates of internal noise $\hat{\sigma}_n$ were computed using the same settings of linear regression as above.

First, we compared methods on their capacity to retrieve true internal noise values, for a range of internal noise, in a typical $n = 1000$ trial experiment. To do so, we let zero-bias, ideal observers with internal noise values $\in [0, 5]$ encounter simulated single- ($n = 1000$ random trials) and double-pass ($n = 500$ random trials + $n = 500$ repeated) experiments, and computed internal noise estimates over 10 runs. For a fair comparison, Double-pass methods were estimated on experiments with the same total number of trials as the other 3 methods, i.e., 500 repeated trials vs 1000 single-pass trials. We then compared variants on the basis of RMSE to the first diagonal, corresponding to true internal noise values.

Second, we investigated how the precision of the methods depended on the number of experimental trials, in particular for shorter experiments with $n < 1000$ trials. We let zero-bias, ideal observers with internal noise values $\in [0, 5]$ encounter single- and double-pass experiments with number of trials $\in [100, 1000]$. For each number of trials, we estimated internal noise, computed the absolute relative error compared to the true value, and averaged it over the range of internal noise values.

4 Results

4.1 Evaluation of Intercept design parameters

First, we evaluated the impact of how many bins are used to discretize trial-pair distances, in both the poynomial and GLM fitting methods, using the criteria of RMS error to the traditional souble-Pass estimation of probability of agreement. Both fitting methods reached their minimum RMSE around 100 bins, with no performance gain observed beyond that point (Figure4).

Then, we compared the performance of binned variants with this optimal number (100) to variants of the method that did not use any binning, for both fitting methods. The lowest RMSE (0.014) was achieved with GLM fitting and no binning, compared to 0.033 for the best polynomial configuration with binning (Fig. 5. GLM fitting without binning provided more accurate $\hat{p}_{agree}$ estimates at higher levels of internal noise (>1), but tended to underestimate the baseline double-pass $\hat{p}_{agree}$ at low internal noise. Conversely, polynomial fitting without binning performed better at lower internal noise levels (<1), but underestimated the double-pass $\hat{p}_{agree}$ at large internal noise.

In the following, we therefore retain the unbinned GLM variant of the Intercept method as optimal for comparison with the other methods (Section 4.4, below).

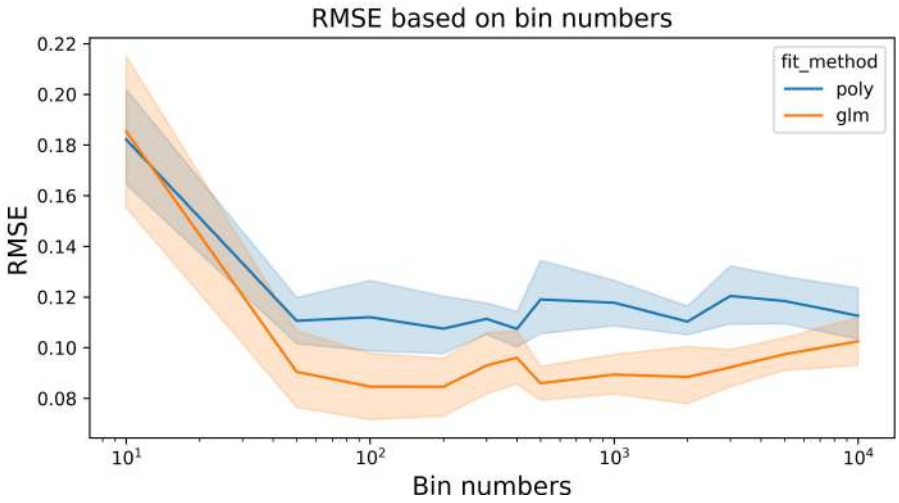


Fig. 4 Influence of number of bins used to estimate the probability of agreement using the Intercept method. Trial-pair distances from simulated experiments were discretized into 10 to 10,000 bins, before being fitted with a polynomial (blue) or GLM (orange). The intercept of the fitted curve was compared to the double-pass estimate of probabiliy of agreement across a range of internal noise values using the root-mean square error (RMSE) metric. 95% confidence intervals computed across 10 independent runs of the same simulation settings.

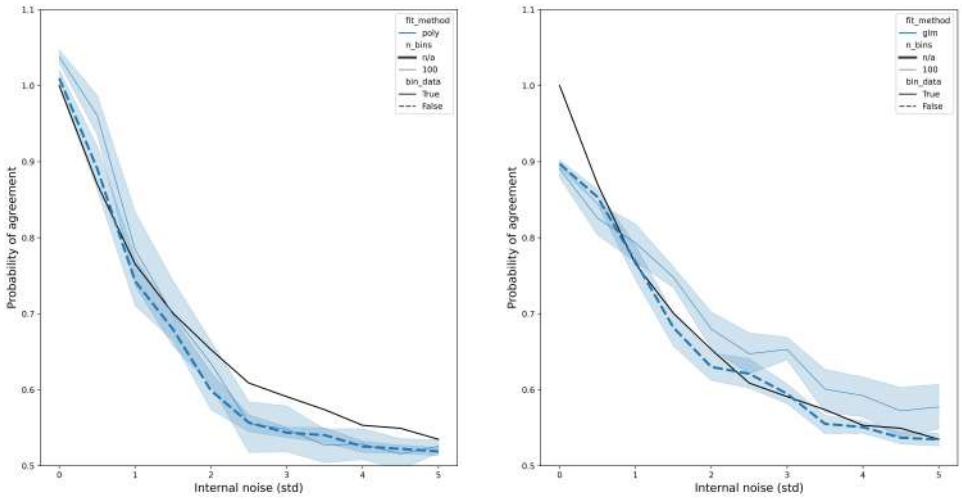


Fig. 5 Estimated probability of agreement over a range of internal noise values, across binning and fitting procedure variants of the Intercept method. Left: Comparison of 100-bins (solid blue) and non-binning (dashed blue) estimations over simulated data for the polynomial fitting variant. Right: Comparison of 100-bins (solid blue) and non-binning (dashed blue) estimation over simulated data for the GLM fitting variant. Solid black curve in both panels mark the double-pass estimate of $\hat{p}_{agree}$ over the same data. 95% confidence intervals computed across 10 independent runs of the same simulation settings.

4.2 Evaluation of Accuracy design parameters

We evaluated the Accuracy estimate $\hat{p}_{agree}$ over three different trial masks (all trials, hits, CRs) and two weighting schemes (weighted vs unweighted). The method produced significantly higher estimates of $\hat{p}_{agree}$ than the double-Pass groundtruth (Fig.6), with trial weighting generally produced higher estimates than unweighted versions. The lowest RMSE (0.311) was obtained with the all-trials, unweighted configuration, with no clear difference from the hit and CR unweighted variants.

While the method clearly does not produce estimates on the same scale as double-pass $\hat{p}_{agree}$ (compare with Intercept, above), it still provides a proxy with a monotonous relation to the double-pass estimate, and in quasi-linear relation to the underlying internal noise. We therefore retain both all-trial variants of the method (weighted, and non weighted) in the following comparison, by using specific lookup table converting these two estimates to internal noise.

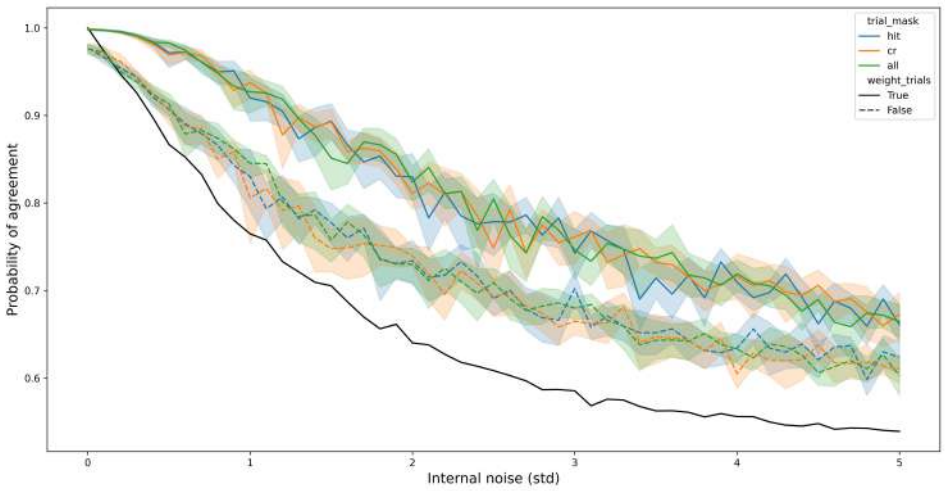


Fig. 6 Estimated probability of agreement across internal noise levels for three different trial masks (all trials, hits, CRs) and two weighting schemes (weighted vs unweighted) of the Accuracy method. Solid lines correspond to trial-weighted conditions (all trials, hit trials, and correct rejections), while dashed lines represent their unweighted counterparts. The black curve indicates the baseline $\hat{p}_{agree}$ values obtained from the double-pass method. 95% confidence intervals computed across 10 independent runs of the same simulation settings.

4.3 Evaluation of GLM-method design parameters

We first evaluated the influence of jitter level on the GLM method's evaluation of internal noise $\hat{\sigma}_n$, with fixed aggregation mode (`argmax`) and link function (`logit`). As expected, increasing jitter had a noticeable mitigation influence on the overestimation of internal noise for small noises (RMSE for internal noise ¡ 1: jitter 10%: 0.70, 15%: 0.74, jitter 1%: 1.89, jitter 2%: 1.37), but also decreased the quality of estimation for large internal noise (RMSE over all σ_n: jitter 10%: 1.20, 15%: 1.42, jitter 1%: 1.09, jitter 2%: 0.93), see Fig. 7. The best RMSE (0.93) was achieved for 2% jitter.

Next, we compared variants across aggregation modes (`argmax`,`argmedian`,`median`,`min`) and link function choices (`logit`,`probit`), while fixing jitter at its optimal 2% value. Aggregation mode had a larger effect than link function (Fig. 8), with `argmax` and `min` achieving comparable performance regardless of link function (`argmax-probit`:RMSE 0.90; `argmax-logit`: 0.93; `min-probit`: 1.00, `min-logit`: 1.03).

In the following, we therefore retain as optimal both the `argmax` and `min` variants, with `probit` and 2% jitter.

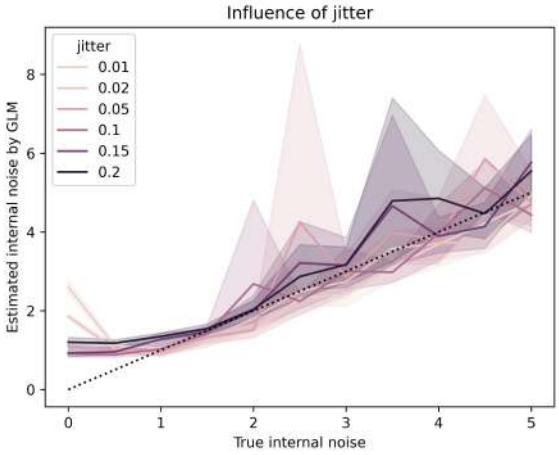


Fig. 7 Estimated internal noise for different jittered variants of the GLM method, across a range of true simulated internal noise values. Dashed line marks the first diagonal (1:1 estimation). 95% confidence intervals across 10 independent runs of the same simulation settings.

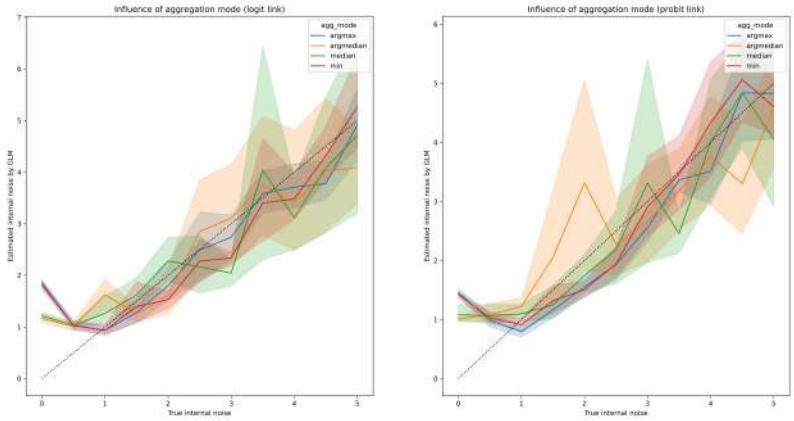


Fig. 8 Estimated internal noise for different aggregation modes and link function variants of the GLM method, across a range of true simulated internal noise values. Left: `logit` estimations. Right: `probit` estimations. Dashed line marks the first diagonal (1:1 estimation). 95% confidence intervals across 10 independent runs of the same simulation settings.

4.4 Comparison of internal noise estimations across methods

We then compared the optimal variant of each method (Intercept: GLM, no binning; Accuracy: all trials, weighted and non-weighted; GLM: 2% jitter, argmax/min, probit) against each other and against the double-pass method, over their capacity to estimate true internal noise for simulated experiments of $n = 1000$ trials. By computing the relative error (in % of internal noise) at $n = 1000$ trials (Fig 9 -green) shows the estimated internal noise against the true internal noise across 100, 600, and 1000 trials, both the Intercept (RE:23%) and Accuracy (RE:25%) methods provided better estimates than the traditional double-pass method (RE:33%). The GLM method (RE:38%) was less accurate as double-pass, but mostly because of its overestimation of low internal noise values. When restricted to $\sigma_n > 1$, GLM was more accurate than double-pass (RE:19%), and in fact both other methods. Fig 9 shows the estimated internal noise against the true internal noise across 100, 600, and 1000 trials, with dashed lines indicating ±30% relative error margins around the diagonal.

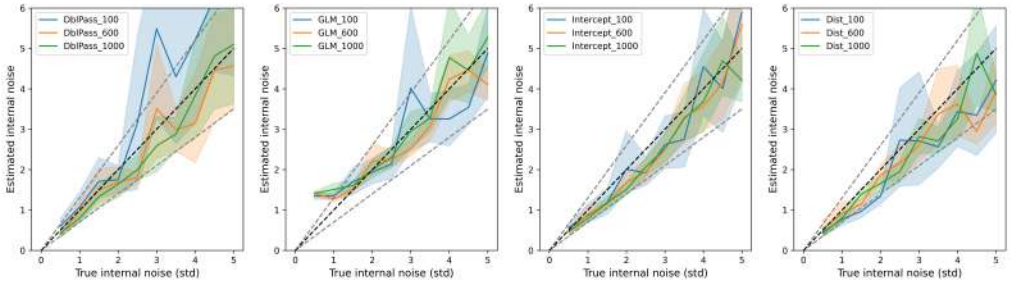


Fig. 9 Estimation accuracy of internal noise across methods and trial counts Each panel shows the estimated internal noise as a function of true internal noise for all four methods: (left-to-right) `Double_pass`, `GLM`, `Intercept` and `Accuracy` across three trial counts (100, 600, 1000). Solid lines represent the estimated values across simulations. The dashed diagonal represents perfect estimation, while outer lines indicate ±30% error margins. Estimation accuracy improves with more trials, and both the Accuracy and Intercept methods outperform the traditional double-pass in terms of relative error in each amount of trial.

Finally, we investigated how the precision of the methods depended on the number of experimental trials based on the relative error (in % of internal noise) over a range of internal noise values in simulated experiments ranging from 100 to 10,000 trials (Fig. 10). The ordering of methods seen at n=1000 was preserved over all range of experiments, with Intercept and Accuracy methods consistently reaching better accuracy than double pass. Figure 10 shows that The Double-Pass method has the highest error among all methods up to 1000 trials, but its accuracy improves beyond this point and eventually reaches the same performance level (30% error) as the Accuracy-based method around 10,000 trials. Due to computational constraints (the number of trial pairs increasing exponentionally with n), the Intercept method was only evaluated up to 1000 trials. The GLM method, while showing the highest error overall, mainly due to overestimating low internal noise values, performs best when the analysis is restricted

to internal noise levels >1, outperforming (RE: 19%) all other methods in that range. At low/very-low number of trials (ex. $n = 100$ single-pass, or $n = 50$ repeated trials, as sometimes used with patients - Adl Zarrabi et al. (2024)), the double-pass method evaluates with an unimpressive 48% relative error ([0.8−1.8] around the typical health-participant estimate of 1.3 - Neri (2010)), while the best method (Accuracy) achieves an error of 30%. Error rates below 20% are reached as early as 600 single-pass trials, using the GLM method under the assumption that $\sigma_n > 1$.

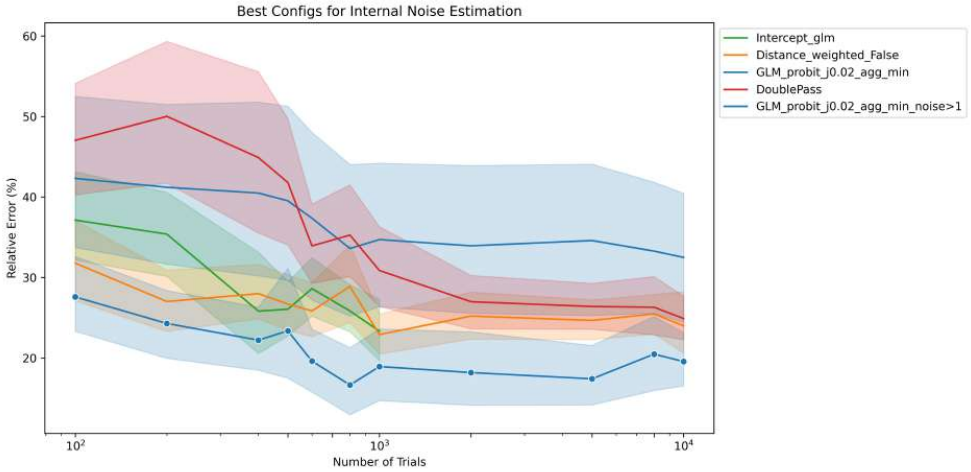


Fig. 10 Relative error of internal noise estimation across methods and trial counts [100, 10000] in log scale. Intercept (`Intercept_glm`) and Accuracy (`Distance_not_weighted`) show consistently lower error than `Double_pass` across all trials counted up to 1000. While `GLM_probit_aggmin` overestimates internal noise at low levels, but it performs best amoung all methods when the noise level is restricted to σ_n >1.

5 Discussion

While one of the important methodological interests of data-driven reverse-correlation experiments is to allow estimating internal noise, the dominant strategy to do so, double-pass, is plagued with important practical limitations. In this work, we introduced and evaluated 3 simple alternative analytical techniques (Intercept, Accuracy and GLM) to estimate internal noise from data-driven experiments in the absence of double-pass measurements. Using computer simulations, we showed that the three techniques consistently outperform double-pass estimation for experiments with low number of trials (<1000), often more than halving percentage error.

The first method, Intercept, provides a heuristic to estimate the double-pass probability of agreement by extrapolating the agreement of pairs of imperfectly repeated trials over the single-pass experiment. GLM fitting with raw binary agreements provided the best agreement with double-pass estimates of probability of agreement,

notably for larger levels of internal noise. The second method, Accuracy, estimates a proxy of p_{agree} based on the observer's agreement with an ideal observer over the same data. The version without trial weighting achieves better performance with no clear difference between unweighted variants of the hit and CR trial masks. Both methods suggest that information about consistency can be extracted from trial and response variability in single-trial data, without requiring the availability of double-pass data. This is important because, in typical experiments, single-pass trials are more numerous than double-pass trials, which typically represent a more restricted set, so there exists a compromise between the accuracy of double-pass estimation of internal noise, and single-pass estimation of kernel. By capitalizing on single-trial trials to estimate p_{agree}, the Intercept and Accuracy method allow avoiding this compromise.

The third method, GLM, is a model-based approach that estimates internal noise as a function of normalized confidence interval around GLM kernel weights. Among all combinations, the `argmax` and `min` showed the best performance when using 2% jitter, and the difference between the logit and probit link functions was minimal. By providing both kernel and internal noise estimates from the same model, the GLM model provides a principled way to analyse data-driven results and, as the other methods, allow to capitalize on more numerous single-pass trials without requiring double-pass data. Compared to Intercept and Accuracy, however, the GLM method suffers from poor (over-estimated) estimates of internal noise at low levels ($\sigma_n < 1$) and low n trials, because of the difficulty to estimate confidence intervals with over-separated training data. This makes the method more appropriate for experimental contexts where participants are expected to have moderate to large levels of $\hat{\sigma}_n$, i.e. for providing a conservative estimate of patients vs controls.

To evaluate the efficiency of optimal variant of each method, we compared them against each other by the varying number of trials, and assessed their ability to estimate true internal noise. Both the Intercept and Accuracy methods showed the lowest relative error (in % of internal noise), followed by GLM, with all three outperforming the Double-pass method up to 1000 trials. As the number of trials increased up to 10000, estimation error decreased (30% error) for all methods. Notably, when the GLM method was filtered to include only internal noise values greater than 1, it became the most accurate method, reaching the lowest error of 19%. These results suggest that double-pass estimation is almost entirely inappropriate for all experiments with $n < 1000$ trials, but that such experiments can be analysed with methods such as Intercept, Accuracy or, in situations where internal noise can be assumed to be > 1, GLM.

One limitation of the present study is that we explored method accuracy over a range of true internal noise values, but did not explicitly examine the effect of response bias b. While b is explicitely accounted for in the lookup tables (based on the proxy of p_{first}, we assumed $b = 0$ in most simulations. It is possible that some method are more robust to bias than others. It is also possible that a direct model-based estimate of bias can be extracted from a GLM model, e.g. via the proxy of the GLM intercept (Murray, 2011). In future work, we aim to investigate how bias can be estimated directly and whether it influences the robustness of internal noise estimation in direct models.

Another limitation lies in the use of a lookup-based backward models to convert $\hat{p}_{agree}$ estimates into $\hat{\sigma}_n$. Lookup tables are impractical, because they requires training a separate model for each estimation method. They are also intrinsically limited in precision due to the fixed search grid (fixed here to 0.1 step size, i.e. a 10% RE at $\sigma_n = 1$). Because the analytical relation between p_{agree} and σ_n is non-linear (Appendix A), lookup search are also associated with saturation effects at large σ_n values, where the nearest grid point typically underestimates the true σ_n. While lookup tables are the (oft, undocumented) default approach in the literature (Goupil et al., 2021; Neri, 2010; Ponsot et al., 2018), it is possible that optimization-based approaches (i.e. with no fixed grid) could provide more accurate parameter estimation, and that these could be combined with the present methods for even better estimates.

Finally, a major limitation of double-pass design is their assumption of stationary noise throughout the experiment. Because the new methods of Accuracy, Intercept and GLM do not rely on the availability of repeated trials, they offer the possibility to estimate internal noise continuously over time, and empirically test whether this assumption is correct. In the literature, no prior study has explored this dynamic aspect of internal noise. As research in cognitive science increasingly focuses on dynamic and adaptive task behavior (Li, Shi, Li, & Collins, 2024; Sergent et al., 2021), being able to track internal noise across time could reveal new insights. For instance, experiments using a block-based or mixed design of template matching could benefit from estimating internal noise within each block to explore the temporal evolution of internal noise.

All in all, with double-pass errors in the $[30\% - 50\%]$ range, this may call for a re/analysis of the published literature where double-pass internal noise estimates were calculated with small number of trials, e.g. in the case of fatigable patients (Adl Zarrabi et al., 2024) or online participants. In studies investigating internal noise and its link to individual traits (e.g., autistic characteristics - Merchie et al. (2024))the lack of observed correlation may partly arise from the imprecision of the traditional double-pass method. The PALIN open-source framework, available both in Python and R, provides all the methods mentioned above that can be applied to reanalyze such datasets, allowing direct comparison between estimation methods.

Declarations

Author contribution. JJA, AAZ, LN and MV, contributed to the conception and design of the study. JJA and MV provided the concept for the Intercept method, LN for the Accuracy method and AAZ for the GLM method. LN provided the concept for the analytical solution of Appendix A. JJA and AAZ wrote the code of the tool in Python and LN wrote the code in R. AAZ and JJA produced the figures and drafted the manuscript, with contributions from LN and MV.

Funding. This work was supported by grant from Fondation pour l'Audition (FPA RD 2021-12).

Code availability. All simulations are done by PALIN toolbox, available in Python and R language, v2.0.0 available at https://github.com/neuro-team-femto/palin.

Conflicts of interest/Competing interests. No author of this work has a direct or indirect confict of interest related to the publication of this work.

Ethics approval and consent to participate. Not applicable.

Consent to publication. Not applicable.

Appendix A Semi-analytical solution for the generative (forward) SDT model

The SDT model assumes that an observer makes a choice based on the noisy internal representations of the stimuli. The internal decision variable is modelled as:

$$X = s_i + \sigma_{ir} \tag{A1}$$

where: $s_i \sim \mathcal{N}(0,1)$ represents the difference between stimuli in the two intervals i and $\sigma_{ir} \sim \mathcal{N}(0,\sigma^2)$ is the internal noise added independently to each repetition r. The observer chooses stimulus 1 if:

$$s_i + \sigma_{ir} > c \tag{A2}$$

where c is the response bias. Thus, the probability of choosing stimulus 1 (first interval) is:

$$\Pr_{\text{first}} = \Pr(s_i + \sigma_{ir} > c) \tag{A3}$$

Since both s_i and σ_{ir} are normally distributed, their sum follows:

$$s_i + \sigma_{ir} \sim \mathcal{N}(0, 1+\sigma^2) \tag{A4}$$

Rewriting the probability:

$$\Pr_{\text{first}} = \Pr\left(\frac{s_i + \sigma_{ir}}{\sqrt{1+\sigma^2}} > \frac{c}{\sqrt{1+\sigma^2}}\right) \tag{A5}$$

Using the standard normal cumulative distribution function (CDF), $\Phi(x)$, we obtain:

$$\Pr_{\text{first}} = 1 - \Phi\left(\frac{c}{\sqrt{1+\sigma^2}}\right) \tag{A6}$$

Agreement occurs when an observer gives the same response in two independent presentations of the same stimulus. The observer makes two independent decisions:

$$X_1 = s_i + \sigma_{i1}, \quad X_2 = s_i + \sigma_{i2} \tag{A7}$$

where $\sigma_{i1}, \sigma_{i2} \sim \mathcal{N}(0,\sigma^2)$ are independent samples of internal noise. Agreement occurs if both repetitions yield the same choice:

$$(s_i + \sigma_{i1} > c \text{ and } s_i + \sigma_{i2} > c) \quad \text{or} \quad (s_i + \sigma_{i1} < c \text{ and } s_i + \sigma_{i2} < c) \tag{A8}$$

The probability of choosing the first stimulus in a single trial is:

$$\Pr_{\text{stim1}} = 1 - \Phi\left(\frac{c-s}{\sqrt{1+\sigma^2}}\right) \tag{A9}$$

The probability of choosing the second stimulus is:

$$\Pr_{\text{stim2}} = \Phi\left(\frac{c-s}{\sqrt{1+\sigma^2}}\right) \tag{A10}$$

Thus, the average (expected) probability of agreement is obtained by integrating over all possible s values:

$$\begin{aligned}\Pr_{\text{agree}} &= \int_{-\infty}^{\infty}\left[\left(1-\Phi\left(\frac{c-s}{\sigma}\right)\right)^2 + \Phi\left(\frac{c-s}{\sigma}\right)^2\right]\phi(s)ds \\ &= \mathbb{E}_{s\sim\mathcal{N}(0,1)}\left[\Phi\left(\frac{s-c}{\sigma}\right)^2 + \left(1-\Phi\left(\frac{s-c}{\sigma}\right)\right)^2\right]\end{aligned} \tag{A11}$$

where $\phi(s)$ is the standard normal density (PDF) of s:

$$\phi(s) = \frac{1}{\sqrt{2\pi}}e^{-s^2/2} \tag{A12}$$

This integral does not have a closed-form solution, but can be approximated using numerical methods (as implemented in the `palin` package). This solution highlights a few key properties of the SDT model. Namely, if there is no internal noise (i.e., $\sigma = 0$), agreement becomes deterministic:

$$\Pr_{\text{agree}} = \int_{-\infty}^{\infty}\left[\mathbb{I}(s>c)^2 + \mathbb{I}(s<c)^2\right]\phi(s)ds = 1 \tag{A13}$$

where $\mathbb{I}(\cdot)$ is the indicator function. When there is no bias (i.e., $c = 0$), the decision rule is symmetric, ensuring that $\Pr_{\text{agree}}$ depends only on σ.

References

Adl Zarrabi, A., Jeulin, M., Bardet, P., Commère, P., Naccache, L., Aucouturier, J.-J., ... Villain, M. (2024). A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after right-hemisphere stroke. *Scientific Reports*, *14*(1), 15194,

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., Damasio, A.R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, *433*(7021), 68–72,

Adolphs, R., Nummenmaa, L., Todorov, A., Haxby, J.V. (2016). Data-driven approaches in the investigation of social perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150367,

Ahumada, A., & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751–1756,

Allard, R., & Faubert, J. (2006). Same calculation efficiency but different internal noise for luminance-and contrast-modulated stimuli detection. *Journal of Vision*, *6*(4), 3–3,

Burgess, A.E., & Colborne, B. (1988). Visual signal detection. iv. observer inconsistency. *Journal of the Optical Society of America A*, *5*(4), 617–627,

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571,

Eggermont, J., Johannesma, P., Aertsen, A. (1983). Reverse-correlation methods in auditory research. *Quarterly reviews of biophysics*, *16*(3), 341–414,

Faisal, A.A., Selen, L.P., Wolpert, D.M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, *9*(4), 292–303,

Gosselin, F., & Schyns, P.G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, *41*(17), 2261–2271,

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, *12*(1), 861,

Hasan, B.A.S., Joosten, E., Neri, P. (2012). Estimation of internal noise using double passes: Does it matter how the second pass is delivered? *Vision research*, *69*, 1–9,

Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., Schyns, P.G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244,

Knoblauch, K., & Maloney, L.T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16), 10–10,

Lee, Y., & Schetzen, M. (1965). Measurement of the wiener kernels of a non-linear system by cross-correlation. *International Journal of Control*, *2*(3), 237–254,

Li, J.-J., Shi, C., Li, L., Collins, A.G. (2024). Dynamic noise estimation: A generalized method for modeling noise fluctuations in decision-making. *Journal of Mathematical Psychology*, *119*, 102842,

Lu, Z.-L., & Dosher, B.A. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychological review*, *115*(1), 44,

Mangini, M.C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226,

Marmarelis, P.Z., & Naka, K.-I. (1972). White-noise analysis of a neuron chain: an application of the wiener theory. *Science*, *175*(4027), 1276–1278,

Merchie, A., Ranty, Z., Adl Zarrabi, A., Bonnet-Brilhault, F., Houy-Durand, E., Aucouturier, J., Gomot, M. (2024). Intact representation of vocal smile in autism: A reverse correlation approach. *PsyArxiv*, 30,

Murray, R.F. (2011). Classification images: A review. *Journal of vision*, *11*(5), 2–2,

Neri, P. (2010). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review*, *17*(6), 802–808,

Neri, P., & Heeger, D.J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature neuroscience*, *5*(8), 812–816,

Owen Brimijoin, W., Akeroyd, M.A., Tilbury, E., Porr, B. (2013). The internal representation of vowel spectra investigated using behavioral response-triggered averaging. *The Journal of the Acoustical Society of America*, *133*(2), EL118–EL122,

Park, W.J., Schauder, K.B., Zhang, R., Bennetto, L., Tadin, D. (2017). High internal noise and poor external noise filtering characterize perception in autism spectrum disorder. *Scientific reports*, *7*(1), 17584,

Pelli, D.G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, *2*(9), 1508–1532,

Ponsot, E., Burred, J.J., Belin, P., Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, *115*(15), 3972–3977,

Ponsot, E., Susini, P., Saint Pierre, G., Meunier, S. (2013). Temporal loudness weights for sounds with increasing and decreasing intensity profiles. *The Journal of the Acoustical Society of America*, *134*(4), EL321–EL326,

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166,

Russell, D.F., Wilkens, L.A., Moss, F. (1999). Use of behavioural stochastic resonance by paddle fish for feeding. *Nature*, *402*(6759), 291–294,

Sergent, C., Corazzol, M., Labouret, G., Stockart, F., Wexler, M., King, J.-R., ... Pressnitzer, D. (2021). Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nature communications*, *12*(1), 1149,

Solomon, J.A., & Pelli, D.G. (1994). The visual filter mediating letter identification. *Nature*, *369*(6479), 395–397,

Thoret, E., Andrillon, T., Gauriau, C., Leger, D., Pressnitzer, D. (2024). Sleep deprivation detected by voice analysis. *PLoS computational biology*, *20*(2), e1011849,

Varnet, L., Wang, T., Peter, C., Meunier, F., Hoen, M. (2015). How musical expertise shapes speech perception: evidence from auditory classification images. *Scientific reports*, *5*(1), 14489,

Venezia, J.H., Hickok, G., Richards, V.M. (2016). Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility. *The Journal of the Acoustical Society of America*, *140*(2), 1072–1088,

Vilidaite, G., & Baker, D.H. (2017). Individual differences in internal noise are consistent across two measurement techniques. *Vision Research*, *141*, 30–39,

Vilidaite, G., & Baker, D.H. (2018). Psychophysical measurement of the effects and non-effects of tms on contrast perception. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *11*(4), 956–957,

Vilidaite, G., Yu, M., Baker, D.H. (2017). Internal noise estimates correlate with autistic traits. *Autism Research*, *10*(8), 1384–1391,

Yan, F.-F., Hou, F., Lu, H., Yang, J., Chen, L., Wu, Y., ... Huang, C.-B. (2020). Aging affects gain and internal noise in the visual system. *Scientific reports*, *10*(1), 6768,

CHAPTER 7

Joint estimation of perseverations and reverse-correlation parameters with the GLM-HMM model

As discussed in Chapter 5, the fact that stroke patients tend to perseverate disrupts both the stimulus-response relation and response variability, leading to large estimation errors in both kernel and noise. First, when patients perseverate for an important proportion of an experiment, kernels estimated with the classification-image or GLM method (Chapter 3) are trying to infer a stimulus-response relation based on data which does not reflect such a decision strategy. Second, if patients perseverate during one or both double-pass blocks, the probability of agreement may be severely under- or over-estimated, leading to internal noise estimates which do not have a direct relation with the patient's perceptual decision process.

One approach to detecting perseveration in behavioral responses is to only analyze choice history, i.e., measure the tendency to repeat the same response across an arbitrary number of consecutive trials (e.g., 15) without considering whether the response

was actually appropriate to the stimulus. While we used this approach to quantify the impact of perseveration in a post-hoc analysis of the results of Chapter 4, it is likely to both falsely label as perseverated streaks of trials which, by chance, genuinely warranted similar responses and also miss shorter perseverations. At the same time, like internal noise, perseverating episodes is not only a source of "estimation noise" that one wants to mitigate but also a symptom of cognitive deficits after stroke. Detecting it accurately (e.g., estimating a true probability of perseverating) may therefore provide crucial insights into post-stroke cognitive impairments and potentially another candidate for biomarkers that can help their diagnosis, prognosis or therapy.

In this chapter, we introduce a new method to conjointly estimate both linear-observer parameters and perseverating episodes, using a joint model with two latent states (input-output hidden Markov model, or GLM-HMM). We show that this model is able to recover perseverating episodes by taking into account not only repeated responses but also stimuli-response relations and to improve the accuracy of kernel and noise estimates across non-perseverated episodes.

7.1 State of the art

7.1.1 Latent states in human decision-making

The assumption of a single-state decision process, as implied, e.g., by the linear-observer model (Chapter 3), fails to capture the dynamical nature of human – and also non-human animal – behaviour. A series of recent works have indeed demonstrated that human cognition is influenced not only by external task demands (e.g., here, stimuli) but also by unknown latent mental processes that change with time and may be only weakly aligned with external task conditions. In Taghia et al. (2018) for instance, authors use a technique akin to a hidden Markov model (Bayesian switching linear dynamical systems; BSDS) to identify brain states with stable spatiotemporal properties from fMRI (functional magnetic resonance imaging) data and show that their occupancy rates during a visual working-memory task correlate with performance. Similarly, recent work has

attempted to identify mind wandering during repetitive decision tasks, using occasional thought probes, and found that trials where participants reported to be off-task corresponded to distinct brain dynamics (Sergent et al., 2021), and that these could be learnt from stimulus-response patterns in a trial-per-trial basis (Zhang & Kool, 2024). In mice, a recent influential paper has also shown that decision-making strategies comprise a single 'engaged' state, in which decisions relied heavily on the stimuli, and several biased states in which errors frequently occurred (Ashwood et al., 2020).

The similarity of these observations with our problem of detecting perseveration in reverse-correlation response data is striking: in a first modeling approximation, perseverating patients may be thought of as being in one of two states: an 'engaged' state (per Ashwood et al. (2020)) where the linear-observer model applies with its kernel and internal noise, and a 'perseverating' state, where the response is determined by the simple rule of being equal to the previous response.

The investigation of time-varying, context-dependent, hidden states that translate either in behavior or brain activity is a challenging computational problem because changes in brain states can be induced by both external stimuli and latent factors, such as motivation, alertness, fatigue and momentary lapse in attention (Taghia et al., 2018). In more recent work, the problem is redefined as the inference of a *state-space model* (SSM) (Commandeur & Koopman, 2007), typically a partially observed Markov model. SSMs have been applied, from their early use in the 1960s Kalman filter for spacecraft tracking (Kalman, 1960), to their application in animal movement modeling (Patterson, Thomas, Wilcox, Ovaskainen, & Matthiopoulos, 2008), and later in complex behavioral models (Roy, Bak, Akrami, Brody, & Pillow, 2021) for capturing hidden cognitive and decision-making processes.

At their core, SSMs consist of two modeling parts: (1) a process model, which captures how the system evolves over time, and (2) an observation model, which maps observations to hidden states, accounting for measurement noise and indirect observations. By decoupling these components, SSMs effectively handle observational errors and reveal latent cognitive dynamics, making them particularly useful for behavioral and neural modeling.

In this work, we show that a particular kind of SSM, the GLM-HMM model (Gen-

eralized linear model - Hidden Markov model), provides an elegant framework to both infer the engaged and perseverant states, as well as the linear-observer parameters of the former, which we have seen can be integrated in the format of a GLM (kernel: GLM weights, Chapter 3; noise: weight confidence intervals, Chapter 6). In the following, we first introduce the methodology of GLM-HMM models, then describe their application to reverse correlation simulated data, and in Chapter 9 we apply it on clinical data.

7.1.2 The GLM-HMM model

7.1.2.1 The hidden Markov model (HMM)

The hidden Markov model (HMM) is a common formalization of the process model in SSMs (Fahrmeir, Tutz, Fahrmeir, & Tutz, 2001). In an HMM, the system being modeled is assumed to be a Markov process with discrete (i.e., hidden) states z_t. In probability theory, a Markov model is a stochastic model used to model randomly changing systems. It is assumed that future states depend only on the current state and not on the events that occurred before it. The observations may be discrete, $y_t \in \{1, 2, \ldots, N_y\}$, or continuous and assumed to be generated from these latent states.

In more detail, an HMM consists of:

- The initial state distribution $p(z_1 = i) = \pi_i$.
- The transition model is noted as $p(z_t = j|z_{t-1} = i) = A_{ij}$ showing how likely to change from one latent state i to j. The diagonal of this matrix A_{ii} shows the percentage of staying in statez_i.
- The observation model has emission probabilities, $p(y_t|z_t = j)$ which specify how observations y_t are generated given the latent state z_t. For continuous observations, a Gaussian emission model (Xuan, Zhang, & Chai, 2001) is often used: $p(y_t|z_t = j) = N(y_t; \mu_j, \sigma_j^2)$.

Since latent states are not directly observable, parameter estimation is performed using the Expectation-Maximization (EM) algorithm (McLachlan & Krishnan, 2008),

specifically the Baum-Welch algorithm (Sundaram, 2000), which iteratively refines the model parameters by alternating between two steps:

- E-step (estimation): estimates the probability of being in each hidden state at every point in time based on the observations. This is done using the Forward-Backward Algorithm (Yu & Kobayashi, 2003), which calculates the posterior probabilities of hidden states (probability distribution over the hidden states at a given time, given the entire sequence of observed data up to that point).

- M-step (maximization): Using these probabilities, updates the model parameters, including the transition matrix (how hidden states change over time) and the emission parameters (how hidden states affect the observations). The goal is to maximize the likelihood of the observed data.

7.1.2.2 Input-output HMMs

The Input-Output HMM (IO-HMM) was first introduced by Bengio and Frasconi (1996) and extends the classical HMM by allowing external inputs x_t to influence both latent state transitions and emissions. This is unlike standard HMMs, in which the distributions of the output variables are conditioned solely on the states. The transition model is made input-dependent in IO-HMM, $p(z_t = j|z_{t-1} = i, x_t)$ and the observation model is also conditioned on x_t: $p(y_t|z_t = j, x_t)$. This modification allows latent state changes to be influenced by contextual information, making IO-HMMs more flexible for modeling stimulus-driven behaviors.

7.1.2.3 The Generalized Linear Model (GLM) HMM

The Generalized Linear Model Hidden Markov Model (GLM-HMM) extends the HMM by incorporating external inputs x_t into the observation model, which models observations as a function of both latent states and input-dependent covariates (Ashwood, 2022). In the GLM-HMM, the HMM governs the distribution over latent states, while a state-specific GLM specifies the strategy of decision-making within each state. Unlike IO-HMMs, where

latent state transitions are input-dependent, in GLM-HMMs, transitions remain Markovian, while the observations y_t depend on both latent states z_t and external inputs, x_t making the GLM-HMM particularly useful for modeling state-dependent decision-making processes.

In more detail, the GLM maps observations y_t to a weighted combination of covariates (typically stimulus, but also e.g., previous stimuli, previous responses, etc.) through a sigmoidal function, modeling the probability of a binary decision. The probability of a binary response y_t in a given latent state z_t is defined as:

$$p(y_t = 1|x_t, z_t) = \frac{1}{1 + e^{-x_t.w_{z_t}}} = \sigma(x_t.w_{z_t}). \tag{7.1}$$

7.1.3 Training algorithms for the GLM-HMM

As in the classical HMM model, the parameters of the GLM-HMM are estimated using the Expectation-Maximization (EM) algorithm. EM iteratively optimizes the parameters by alternating between computing posterior probabilities of latent states (E-step) and updating model parameters (M-step). Depending on the estimation method, we can use either Maximum Likelihood Estimation (MLE) or Maximum A Posteriori (MAP) Estimation (Dempster, Laird, & Rubin, 1977).

7.1.3.1 Maximum-Likelihood estimation (MLE)

MLE estimates the parameters by maximizing the likelihood of the observed data:

$$\hat{\Theta} = \arg\max_{\Theta} P(y_{1:T}|x_{1:T}, \Theta) \tag{7.2}$$

where $\hat{\Theta}$ represents the optimal parameter estimates based purely on observed data.

Since the latent states are unobserved, the EM algorithm is used to iteratively refine the model parameters. It consists of two steps:

- E-Step: Using the Forward-Backward algorithm, we compute the posterior probability of each latent state:

$$\gamma_{t,k} = P(z_t = k|y_{1:T}, x_{1:T}, \Theta) \tag{7.3}$$

where $\gamma_{t,k}$ represents the probability of being in latent state k at time t, the entire sequence of observations and inputs.

Additionally, we compute the expected state transitions:

$$\xi_{t,ij} = P(z_t = i, z_{t+1} = j | y_{1:T}, x_{1:T}, \Theta) \tag{7.4}$$

which represent the probability of transitioning from state i to state j.

- M-Step: Using the posterior estimates from the E-Step, we update the model parameters by maximizing the expected log-likelihood. The transition matrix A is updated as:

$$A_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{t,ij}}{\sum_{t=1}^{T-1} \gamma_{t,i}} \tag{7.5}$$

ensuring that each row of the matrix represents valid transition probabilities summing to one. The GLM parameters w_k are updated by maximizing the conditional log-likelihood:

$$\max_{w} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{t,k} \log P(y_t | x_t, w_k) \tag{7.6}$$

which is typically optimized numerically using gradient ascent, as no closed-form solution exists for the optimal weights.

7.1.3.2 Maximum A Posteriori (MAP) Estimation

MAP extends MLE by incorporating prior distributions on the model parameters to regularize estimation. It maximizes the posterior probability:

$$\hat{\Theta} = \arg\max_{\Theta} P(y_{1:T} | x_{1:T}, \Theta) P(\Theta) \tag{7.7}$$

where $P(\Theta)$ represents the prior distribution. This ensures that the estimated parameters remain within reasonable ranges, especially when data are limited.

MAP follows the same EM procedure as MLE but with additional regularization from prior distributions. As in MLE, in the E-step we compute the posterior probabilities of the latent states and the expected transition probabilities. What is different here is the M-step, which incorporates prior distributions to regularize parameter updates.

- The transition matrix A is updated similarly to MLE but constrained by a Dirichlet prior which ensures that transition probabilities remain non-negative and sum to one:

$$P(A) = \text{Dirichlet}(A|\alpha) \tag{7.8}$$

- The GLM parameters w_k are updated by maximizing the posterior probability:

$$\max_{w} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{t,k} \log P(y_t|x_t, w_k) + \log P(w_k) \tag{7.9}$$

  where the prior $P(w_k)$ is modeled as a Gaussian distribution:

$$P(w) = \mathcal{N}(w|\mu, \sigma^2) \tag{7.10}$$

MAP estimation is particularly beneficial when the dataset is limited, as it incorporates prior knowledge to improve generalization and prevent overfitting.

7.2 Model development

To model perseveration, we propose to use a GLM-HMM model in which perseverating patients may be thought of as being in one of two states: an 'engaged' state where the linear-observer model applies with its kernel and internal noise and a 'perseverating' state, where the response is determined by the simple rule of being equal to the previous response.

This method has already been applied to mice decision data, showing that mice dynamically switch between an engaged state, where decisions rely on sensory input, and biased states, where errors are more frequent (Ashwood et al., 2020). Incidentally, this study contributed to the development of the SSM (State Space Model) library by Scott Linderman's lab (Statistics Dept, Stanford University `https://github.com/lindermanlab/ssm`), which I use in my analysis. In Ashwood et al. (2020), the authors aimed to determine the number of states that best explain movement-based decision-making in mice by minimizing the log-likelihood across subjects. However here, unlike Ashwood et al. (2020),

where state transitions were inferred from known correct responses, we do not have ground truth for state switching or correct responses, which are undefined in reverse correlation experiments. To overcome this limitation, we use the PALIN toolbox to simulate multiple perseverating observers with known true states (either engaged or perseverating, with a known probability of transition), allowing us to validate the switching algorithm.

Our GLM-HMM model takes as inputs the difference in features between manipulated stimuli in each trial (as in GLM estimations of kernel, see Chapter 3) and an additional covariate: the observer's previous choice. In perseverating state, an observer's response at trial t is identical to their response at trial $t-1$, and this choice is not based on the stimulus representation. Conversely, in engaged state, their choice is influenced by the stimulus and not (or not mostly) by their previous response. Our goal is to determine, for each trial, whether the observer is in the engaged or perseverative state, providing a state-based interpretation of decision-making behavior.

To fit the conditions of the reverse correlation experiment of Chapter 4, we set the number of states to 2: engaged (ENG) and perseverating (PER), and set the number of input dimensions/covariates to 8, including 7 stimulus-related features and 1 choice history. The outputs were modeled as binary responses (first or second interval). In terms of GLM weights, we would expect that, if correctly inferred from data, the GLM corresponding to the engaged (ENG) state should have non-zero weights on the stimulus features (corresponding to the linear-observer kernel, which can have zero weight on certain features if they are not used in the decision, e.g., the beginning of an interrogative utterance that has less weight as seen in Chapter 4) and zero weight on the previous response choice, while the GLM of the perseverating (PER) state should have zero weight on the stimulus feature and unit weight (or, equivalently, one arbitrarily positive weight) on the previous choice (Figure 7.1, Top-right). In terms of transition matrix, a GLM-HMM trained from patient data will have four transition probabilities (a rank-2 matrix), where diagonal terms (probability to stay engaged p_{22}if engaged and probability to stay perseverating p_{11} if perseverating) are expected to dominate, and non-diagonal (p_{12}, p_{21}) elements govern transition behaviour (Figure 7.1, Bottom).

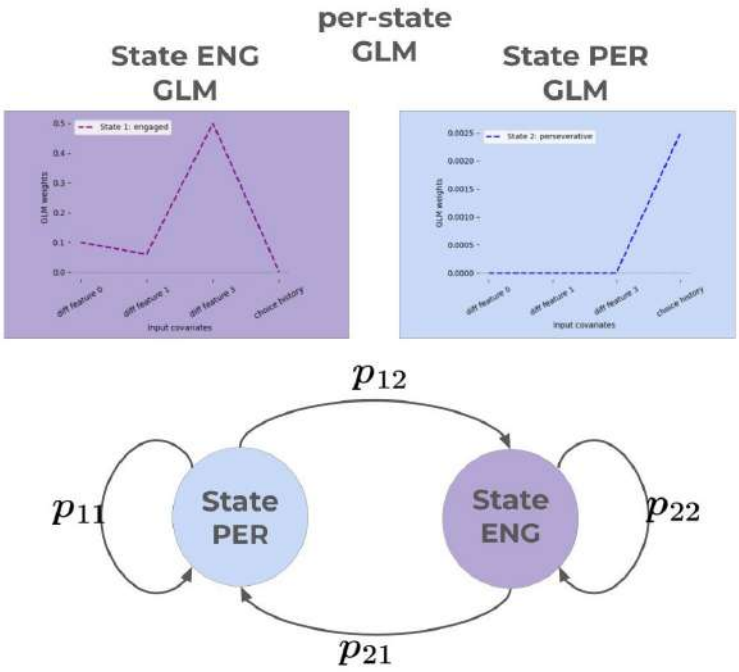


Fig. 7.1 **Per-state GLM weights and state transition structure in the GLM-HMM Top** The estimated GLM weights for each hidden state: the engaged (ENG) state (left, purple) exhibits flexible weights across stimulus features and minimal influence from previous choices, while the perseverative (PER) state (right, blue) demonstrates a dominant weight on previous response and negligible weights on stimulus features. **Bottom** The state transition matrix of the GLM-HMM, where p_{11} and p_{22} represent the probability of remaining in the same state, and p_{12} and p_{21} denote switching between states. Accurate inference of these parameters enables the model to capture dynamic switching between engaged and perseverative behaviors during the task.

7.2.1 Training methods

7.2.1.1 Training with MLE

With Maximum Likelihood Estimation (MLE) with test training data, we consistently observe that 2-state GLM-HMM models do not converge to clearly separated states and, in particular, incorporate non-zero weights on choice history in both states. Since choice history is often predictive of the current choice, the model naturally assigns weights to this covariate in both states, even when one state (the engaged state) should ideally rely solely on stimulus-driven decision-making. This issue arises because MLE does not enforce state-specific constraints, allowing information from past choices to leak into both states rather than being confined to the perseverative state alone.

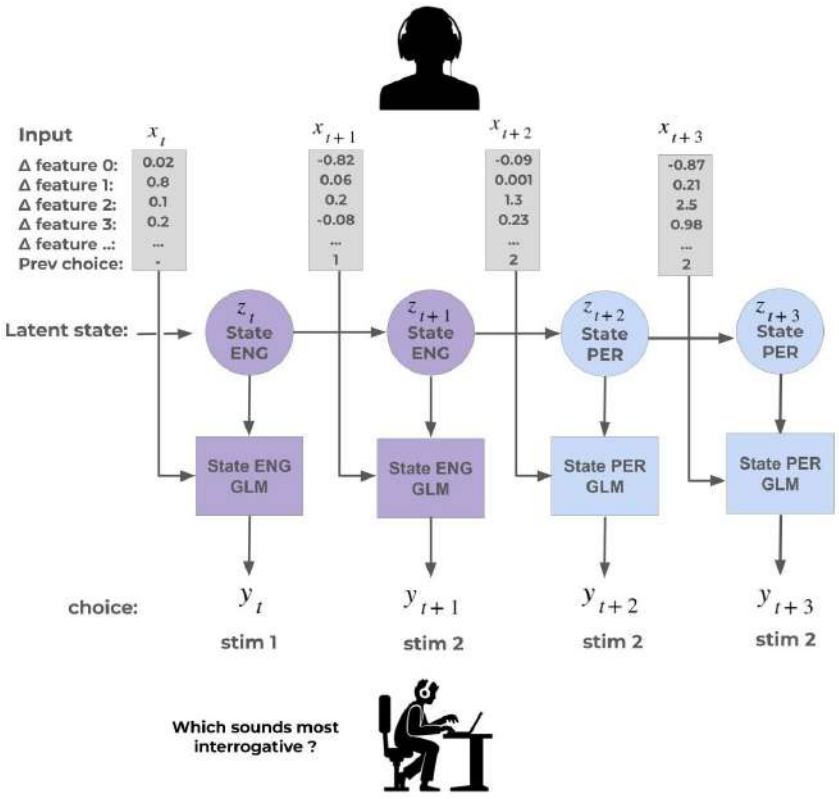


Fig. 7.2 **Illustration of state switching in the GLM-HMM framework** Each trial (t) presents a stimulus characterized by several acoustic features (inputs X). At each time point, the participant is in a latent state z, either engaged (ENG, purple) or perseverative (PER, blue), which determines how the GLM relates the input to the response. State transitions (arrows) can occur dynamically between trials, reflecting fluctuations in cognitive engagement. The output Y is a sequence of choices; in the PER state, the response typically repeats the previous trial's choice

Additionally, MLE is prone to overfitting, capturing any correlation in the data that increases likelihood, even if it does not reflect the true underlying cognitive states. Without a prior to guide the estimation, the model struggles to separate the engaged and perseverative states properly, leading to a failure in state dissociation. Additionally, MLE only seeks to maximize the likelihood of observed choices, which makes it susceptible to local likelihood optima, where the optimization process gets stuck in suboptimal solutions. This is particularly problematic when the dataset is limited in the number of trials.

7.2.1.2 Training with MAP

To overcome these limitations, we used Maximum A Posteriori (MAP) estimation in order to incorporate structured priors that guide the learning process and prevent state ambiguity.

To optimize priors for both GLM weights by Gaussian priors and transition probabilities by Dirichlet priors (Sethuraman, 1994), we designed a systematic validation approach by simulating multiple perseverating observers in PALIN with a known ground truth of state transitions. We used Bayesian optimization (Jones, Schonlau, and Welch (1998), `https://pypi.org/project/bayes-optim/`, (Frazier, 2018)) to explore a broad set of parameter values and identify those that best facilitate the separation of latent states. Unlike grid search or random search, Bayesian Optimization is a sample-efficient method for optimizing costly objective functions. It builds a probabilistic surrogate model to approximate the objective and uses an acquisition function to balance exploration and exploitation, efficiently guiding the search toward optimal hyperparameters.

For the GLM weights, priors were modelled using Gaussian distributions, where the hyperparameters, mean μ and variance σ^2, were optimized to balance flexibility and regularization in state-dependent parameter estimation. The transition probabilities were controlled using a Dirichlet prior with an adjustable concentration parameter α, which dictated the tendency of the model to remain in the same state or transition between states.

The optimization process aimed to find the best combination of hyperparameters that maximized the likelihood of the observed data while preventing overfitting. By iteratively evaluating different prior configurations and updating the surrogate model, Bayesian Optimization efficiently converged toward the best set of priors without the need for exhaustive search.

To operationalize this approach, we developed a pipeline where multiple simulated datasets, each representing a perseverating observer generated via the PALIN framework (see 5.1.4), served as testbeds for prior optimization. Each dataset comprised simulated responses, input stimuli, and the ground-truth sequence of latent states (engaged or per-

severative).

For each simulation, we systematically varied the prior means and variances for the GLM weights in both states, as well as the alpha parameter of the Dirichlet prior governing state transitions. The GLM-HMM was configured with two latent states and eight input features and was fit to the simulated data using Expectation-Maximization with MAP estimation. Specifically, for the engaged state, the prior mean and variance were assigned to the weights corresponding to all stimulus input features, while the weight on the previous choice input was fixed at zero. In contrast, for the perseverative state, the prior mean and variance were applied to the weight on the previous choice input, with all other weights set to zero. This parameterization is reflected in the prior structure: prior_means = [(0, mean_value_1), (mean_value_2, 0)] and prior_sigmas = [(0.01, sigma_value_1), (sigma_value_2, 0.01)], where mean_value_1 and sigma_value_1 specify the prior for the previous choice input in the perseverative state, and mean_value_2 and sigma_value_2 specify the prior for the stimulus features in the engaged state.

The search for optimal priors by Bayesian optimization, iteratively proposed new sets of hyperparameters (means: 0–2, variances: 1–5, alpha: 1–3) and refit the model, minimizing the root mean square error (RMSE) between the inferred and true latent states. RMSE was calculated both globally and for each state separately, providing a sensitive measure of state recovery accuracy.

7.2.2 Measuring goodness of fit

The accuracy of GLM-HMM parameter estimation was evaluated throughout the prior optimization process and subsequent validation (see Section 7.3). For each fitted model, we computed posterior state probabilities and compared the most likely predicted state to the true state at each trial, using RMSE as our primary metric. This was calculated for all states combined and separately for engaged and perseverative states. In addition, we recorded the model's log-likelihood to assess overall fit to the observed data.

Figure 7.3 illustrates the results of GLM-HMM prior optimization for two simulated perseverating observers, showing how different aspects of model fitting and state recovery

are visualized across the panels.

In the top left panel for each observer, the plot depicts the posterior probability of being in the engaged state (green) and the perseverative state (orange) at each trial. These dynamic probabilities are overlaid with the true latent state sequence (gray dashed line), allowing a direct visual comparison of model inference versus ground truth throughout the session.

The top right panel displays the extracted GLM weights for each state, plotted across all eight input features. The engaged state weights are shown in green (right y-axis), while the perseverative state weights are plotted in orange (left y-axis). The x-axis labels the input features: the first seven represent stimulus features, and the eighth corresponds to the previous choice input. For the engaged state, this last weight remains at zero (reflecting no influence of previous response), whereas for the perseverative state, this weight varies and is typically non-zero, capturing the habitual response repetition characteristic of perseveration. The panel titles report the correlation between the estimated and true kernels, as well as with the one-state GLM-HMM kernel for direct quantitative comparison.

In the bottom left panel, showing a comparison of the kernels. The blue line represents the true internal kernel (i.e., the simulated "ground-truth" perceptual weights). The red line shows the kernel estimated by a one-state GLM-HMM, which combines all trials without accounting for state switching. The green line depicts the engaged-state kernel estimated by the two-state GLM-HMM, derived specifically from trials inferred to be in the engaged state. Here, the x-axis covers the seven stimulus feature indices, showing how well each modeling approach captures the true perceptual representation.

The bottom right panel presents the recovered transition matrices for each observer, summarizing the estimated probabilities of remaining in or switching between the engaged and perseverative states. These matrices quantify the model's ability to learn the underlying dynamics of state persistence and transition.

For each observer, details of the best-fitting priors and performance are indicated above the plots. For instance, in the first (top) observer, the optimal priors included a prior mean 1 of 1.44, prior mean 2 of 4.62, prior sigma 1 of 4.09, prior sigma 2 of 2.62, and

a prior alpha of 1.00. This configuration achieved a log-likelihood of −92.34 and a total RMSE of 0.30, with particularly accurate recovery for the engaged state ($RMSE = 0.26$).

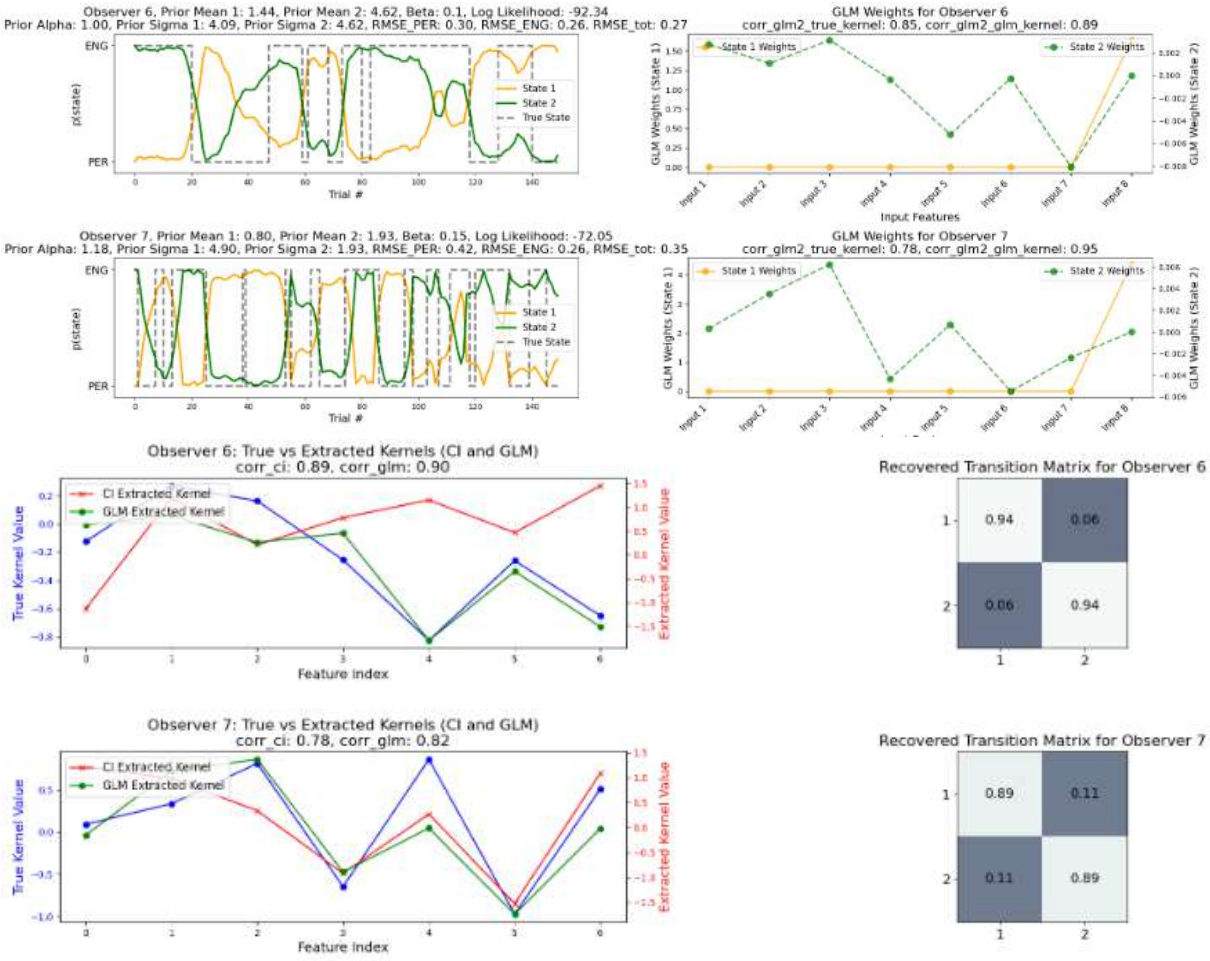


Fig. 7.3 **GLM-HMM prior optimization on simulated perseverating observers Top left:** Posterior state probabilities (green: engaged, orange: perseverative) with the true latent state (gray). **Top right:** State-specific GLM weights across eight input features (stimulus features and previous choice input), with their correlations to the true kernel and single-state GLM-HMM estimate. **Bottom left:** Comparison of true kernel (blue), engaged-state kernel from GLM-HMM (green), and kernel from single-state GLM-HMM (red). **Bottom right:** Recovered transition matrices for each simulated observer.

7.3 Model validation

To evaluate the effectiveness of the selected priors in minimising RMSE, we assess how accurately the model recovers the true latent states in a simulated reverse correlation

experiment with perseverating observers. By comparing the inferred states to the known ground truth, we quantify the extent of estimation errors and determine the reliability of the optimized priors.

7.3.1 RMSE of states

Figure 7.4 illustrates that as the true probability of staying in the perseverative state increases, the model becomes more accurate in predicting perseveration. However, RMSE increases for the engaged state, indicating that the model struggles to correctly identify engaged trials when perseveration dominates. When a significant proportion of trials (around 40%) belong to the engaged state, the model finds it difficult to transition back to engaged once it has classified a sequence as perseverative. However, increasing the number of trials improves overall performance. Ultimately, the total RMSE is primarily driven by the poor fit of the engaged state rather than errors in the perseverative state estimation.

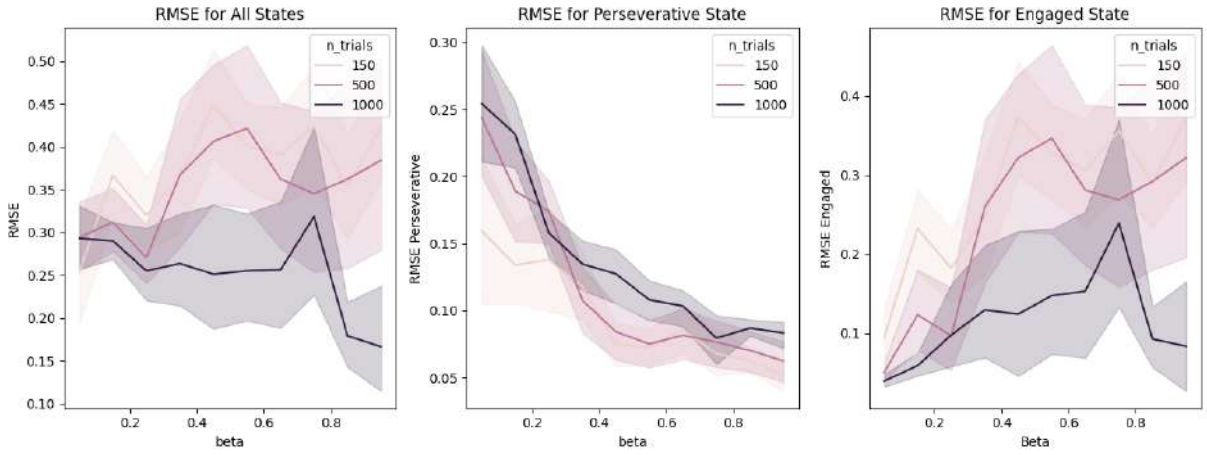


Fig. 7.4 **Root mean square error (RMSE) of state inference as a function of the probability of staying in the perseverative state (β), for different trial counts** The left panel shows total RMSE across both states, the middle panel shows RMSE for the perseverative state, and the right panel shows the engaged state. Increasing β improves detection of perseveration but reduces accuracy for engaged states, especially when trial counts are low.

7.3.2 Precision of kernel in engaged state

Next, we conducted simulations to compare the effectiveness of different methods under varying levels of internal noise and in the presence of a perseverating observer, given their engaged kernel. Both the Classification Image (CI) method and Generalized Linear Model (GLM) estimate the kernel by including possibly perseverated trials. In contrast, the GLM-Hidden Markov Model (GLM-HMM) explicitly identifies perseverated trials and estimates the kernel only for engaged trials. Conversely, when the total number of trials is limited (e.g., 150 trials in the experimental setup of Chapter 4), GLM-HMM may discard a significant portion of data, potentially impacting kernel estimation. Figure 7.5 confirms these predictions.

Unlike CI and GLM, GLM-HMM maintains a slower decline in kernel correlation, indicating that it preserves kernel precision in engaged trials over a broader range of perseveration probabilities (β). At very high perseveration probabilities, GLM-HMM even surpasses the other two methods in kernel accuracy, likely due to its ability to better isolate engaged trials from perseverative noise.

When both perseveration and internal noise increase, GLM-HMM becomes more sensitive in detecting and excluding perseverative trials from the engaged state. This results in a divergence between the estimated engaged-state kernel and the kernel computed over all trials when the perseveration probability exceeds 0.6. As the number of trials increases, the effect of trial exclusion diminishes, allowing GLM-HMM to achieve the same kernel precision as the other two models. Across all methods, kernel precision declines as internal noise levels increase.

7.3.3 Precision of internal noise in engaged state

Finally, we conducted another simulation to estimate internal noise, comparing two methods: the Double-Pass method, which estimates noise across repeated trials, and the GLM-HMM method, which estimates noise in engaged trials using the confidence-interval method introduced in Chapter 6.

Our observations (Figure 7.6) indicate that when the number of trials is limited (e.g.,

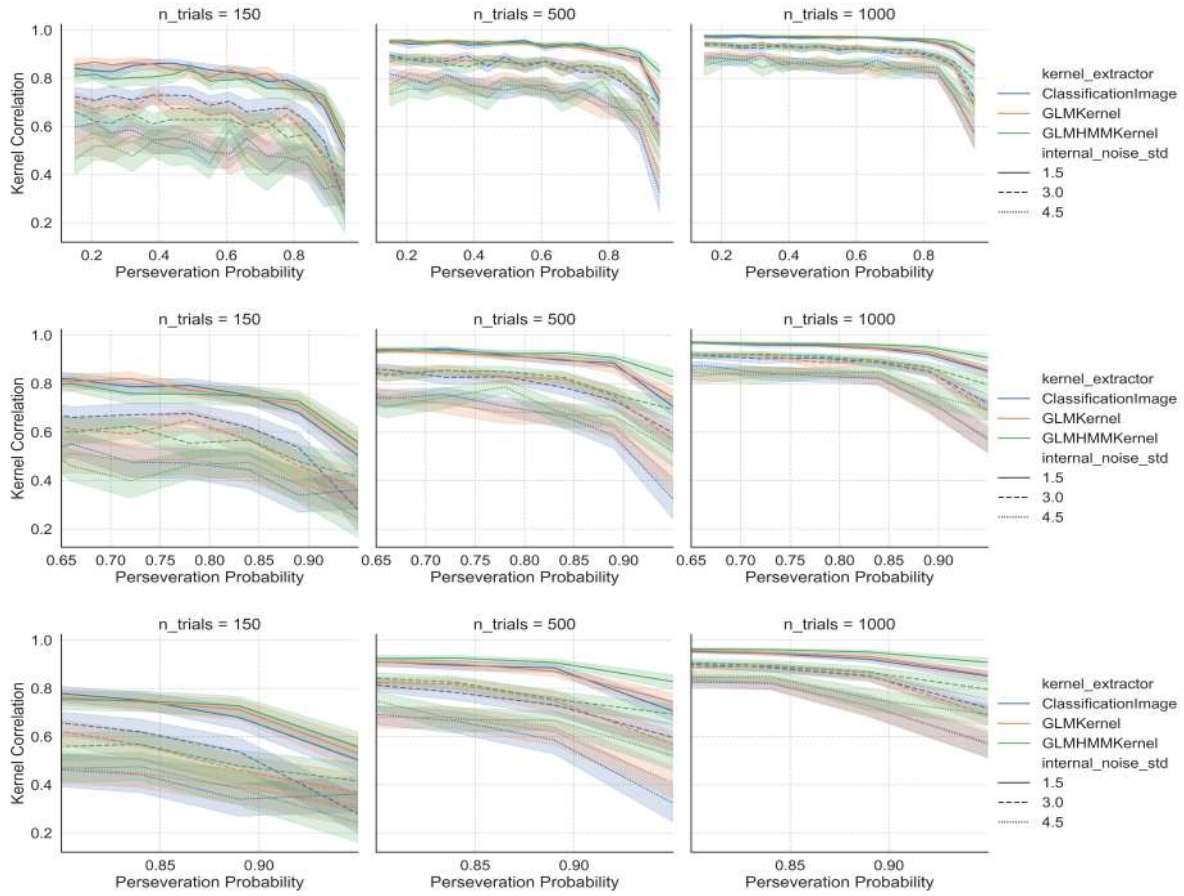


Fig. 7.5 **Effect of perseveration and internal noise on kernel estimation accuracy** Each panel plots kernel correlation for three methods (Classification Image, GLM, GLM-HMM) across different perseveration probabilities (β), for three levels of internal noise ($\sigma_z = 1.5, 3, 4.5$). The top row covers a wide range of perseveration probabilities (0.01 to 0.99), the middle row focuses on high perseveration (0.65 to 0.95), and the bottom row zooms in on very high perseveration (0.85 to 0.95). As perseveration and internal noise increase, GLM-HMM preserves kernel accuracy in the engaged state better than the other methods, especially at high perseveration probabilities.

150 trials) and perseveration is low, both methods perform similarly, with neither showing a clear advantage. However, as perseveration increases (e.g., at a probability of 0.6), the GLM-HMM method provides a more accurate estimation of internal noise, while the Double-Pass method remains relatively unchanged. At very high levels of perseveration, the Double-Pass method struggles to distinguish different noise levels, leading to underestimation. In contrast, the GLM-HMM method remains sensitive to noise variations, suggesting it is better suited for cases where perseveration is a significant factor.

Additionally, as the number of trials increases, the difference between the two methods

becomes even more pronounced. The GLM-HMM method continues to refine its internal noise estimation, while the Double-Pass estimates remain stable but do not improve in distinguishing noise levels, further emphasizing the advantage of GLM-HMM in scenarios with high perseveration and larger datasets.

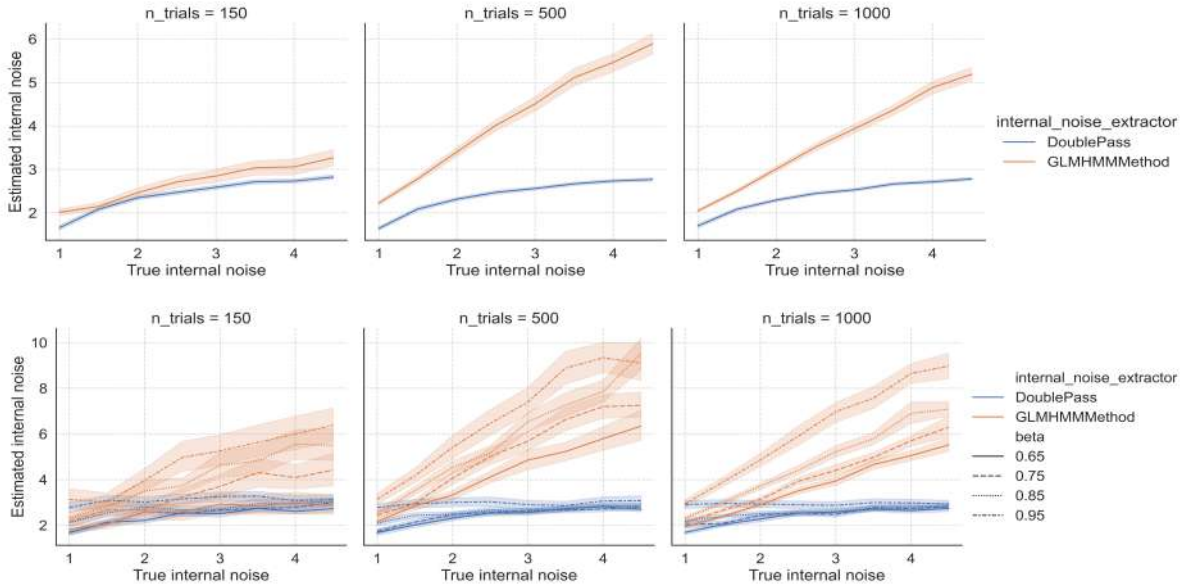


Fig. 7.6 **Internal noise estimation in engaged trials using Double-Pass and GLM-HMM methods Top row:** Estimated internal noise across all perseveration probabilities (β), shown for $n = 150$, 500, and 1000 trials. **Bottom row:** Internal noise estimates are stratified by specific perseveration probabilities ($\beta = 0.65, 0.75, 0.85, 0.95$). GLM-HMM demonstrates improved sensitivity to increasing internal noise, especially at higher perseveration and larger sample sizes.

7.4 Discussion

In Chapters 2 and 3, we gave biological and computational foundations for understanding prosody perception, emphasizing the importance of mechanistic, theory-driven approaches over traditional descriptive assessments. We introduced the reverse correlation method as well as typical approaches via classification images, GLM and double-pass to estimate kernels and internal noise. In Chapter 5, we identified two limitations of these classical methods when confronted with stroke patient data, namely their low robustness to a low number of trials and to perseverating patients. In Chapter 6, we introduced new

internal noise estimation methods that do not require double-pass data. To complement this approach, the current chapter introduced a novel kernel and noise estimation method (the GLM-HMM model) designed to cope with patient perseveration.

By jointly modeling trial-by-trial fluctuations in sustained attention (i.e., transitions between perseverative and engaged states) together with reverse correlation parameters within a single computational framework, our proposed GLM-HMM model enables us to capture the latent temporal dynamics of attention in an ecologically valid way, without interrupting the natural flow of the experiment or relying on subjective self-reports.

By fitting a two-state GLM-HMM to simulated observer data, with each state characterized by distinct GLM weights over stimulus features and response history, we were able to infer periods of engaged versus perseverative responding. The engaged state reflected stimulus-driven, flexible decision-making, whereas the perseverative state was dominated by habitual or history-driven choices, consistent with transient lapses of attention.

Crucially, the model recovers not only latent state transitions but also estimates the stimulus-response kernels and internal noise for each state, yielding a comprehensive, data-driven characterization of attention dynamics and perceptual processing. Our results demonstrate that the proposed joint estimation framework outperforms conventional methods for both kernel and internal noise, particularly in scenarios involving frequent perseveration.

This new model, combined with the confidence-interval method of estimating internal noise, provides a way to reanalyse the patient data of Chapter 4. Not only does the method promise better estimates of reverse-correlation parameters (kernel and internal noise) because they are not plagued with uncertainties linked to perseveration and a low number of trials, but the method also offers the opportunity to analyse patient perseveration parameters (in the form of a state transition matrix). In the next chapter, we will apply this framework to reanalyse the patient data of Chapter 4, and describe how better modeling may preserve or modify the correlations with clinical data already observed with classical methods.

Part IV

A re-analysis of patient data

CHAPTER 8

Re-analysis of psychophysical biomarkers

The main objective of the GLM-HMM model was to apply a probabilistic model to detect hidden states in participants' responses and provide estimates of kernel and internal noise that are not plagued by perseveration.

One of the key findings in Chapter 4 is that both internal representations and internal noise play a major role in differentiating patients from controls, as well as potentially different subtypes of patients. However, we recognized that representations, and perhaps most strikingly, internal noise estimates from the double-pass method were contaminated by perseveration in patients, potentially leading to inflated values. Thus, a primary goal in this chapter is to assess whether removing perseverative trials using the GLM-HMM model would affect kernel and noise estimates in our clinical samples, as well as their correlation with clinical data.

In the following, we first apply the GLM-HMM model to the patient data of Chapter 4 and study how this new model affects estimates of kernel, noise and perseveration probability (section 8.1). We then use these new estimates to re-evaluate whether reverse correlation parameters allow separation of controls and patients (section 8.2), MEC > 9 and MEC < 9 patients (section 8.3), how they correlate with associated clinical measures

(section 8.4), and discuss any potential differences.

8.1 Comparison between new GLM-HMM estimates and old GLM estimates

To fit the GLM-HMM model to patient data, we used the priors estimated by Bayesian optimization on the simulation data in Chapter 7, namely: $\mu_{\text{PER}} = 0.91$, $\mu_{\text{ENG}} = 2.43$, $\sigma_{\text{PER}} = 3.79$, $\sigma_{\text{ENG}} = 2.43$, and $\alpha = 1.82$. We fit the GLM-HMM with these priors to each individual patient and then extracted the posterior probability of each trial, the transition matrix and fitted GLM weights.

8.1.1 Illustrative example

We illustrate the result of this procedure on two example participants, one patient (subj 36) and one control (subj 1).

The fitted patient data (Figure 8.1) exhibit highly fluctuating posterior probabilities, with frequent transitions between the PER and ENG states throughout the session.(Figure 8.27 and Figure 8.28 show these posterior probabilities separately for each RH stroke patient and control participant, respectively.) This instability suggests that the patient does not maintain a stable engagement with the task, potentially reflecting cognitive impairments such as perseveration. The transition matrix confirms this, with a lower probability of remaining in the ENG state ($P_{ENG->ENG} = 0.82$) and a higher likelihood of switching back to PER ($P_{ENG->PER} = 0.18$). The kernel comparison reveals more divergence between the GLM-HMM engaged kernel and the control kernel than the weighted-sum and GLM kernels, indicating that the patient's perceptual processing differs significantly from normative patterns and that classical methods that do not take perseveration into account may obfuscate these differences.

In contrast, fitted data for control participant Subj1 is initially estimated in the PER state for about 40 trials before transitioning fully into the ENG state around trial 40. The transition matrix supports this observation, showing a high probability of staying in either

state once entered ($P_{ENG->ENG} = 0.96$, $P_{ENG->ENG} = 0.99$), indicating stable state dynamics. The kernel weight comparison shows that the GLM-HMM engaged kernel (red) aligns closely with the control kernel (green), suggesting that this subject's perceptual processing follows the normal pattern of controls. The fact that a control participant is estimated as perseverating may either reveal limitations of the GLM-HMM fitting procedure, e.g., in the case where initial responses do not fully conform to the final ENG state kernel, or true fluctuations of response strategies that are not limited solely to pathological situations, possibly reflecting transient cognitive fatigue that shifts behavior toward more habitual, less effortful responses (Pessiglione, Blain, Wiehler, & Naik, 2025). We will return to a more systematic analysis of state transition in controls below.

8.1.2 Impact on kernel estimation

To assess how removing perseverative trials affects kernel estimation, we compared GLM-estimated kernels before and after filtering out perseverating trials, i.e., using the classical GLM kernel vs GLM-HMM kernel estimation.

We first examined whether new estimates modified the proximity of control and patients to the average kernel of the control group (Figure 8.2-left). Among healthy participants, the initial correlation between their GLM kernel and the average control kernel was 0.87, which only slightly decreased to 0.84 after removing perseverative trials. This indicates that perseverative trials introduce some noise, but their removal does not drastically alter kernel estimation in controls. Similarly, in patients, before filtering, the correlation between each patient's GLM kernel and the average control kernel was only 0.22, reflecting a substantial deviation from normative, stimulus-driven responses; after removing perseverative trials, this correlation further dropped to 0.19, suggesting that filtering out these trials does not fully restore a stimulus-driven kernel, and residual distortions remain.

Second, we examined the stability of kernel estimation by comparing the GLM kernel before removal with the GLM-HMM kernel after removal within each group (Figure 8.2-right). In controls, the correlation between these two versions of the kernel was 0.93, indicating that filtering perseverative trials had minimal impact on kernel structure. How-

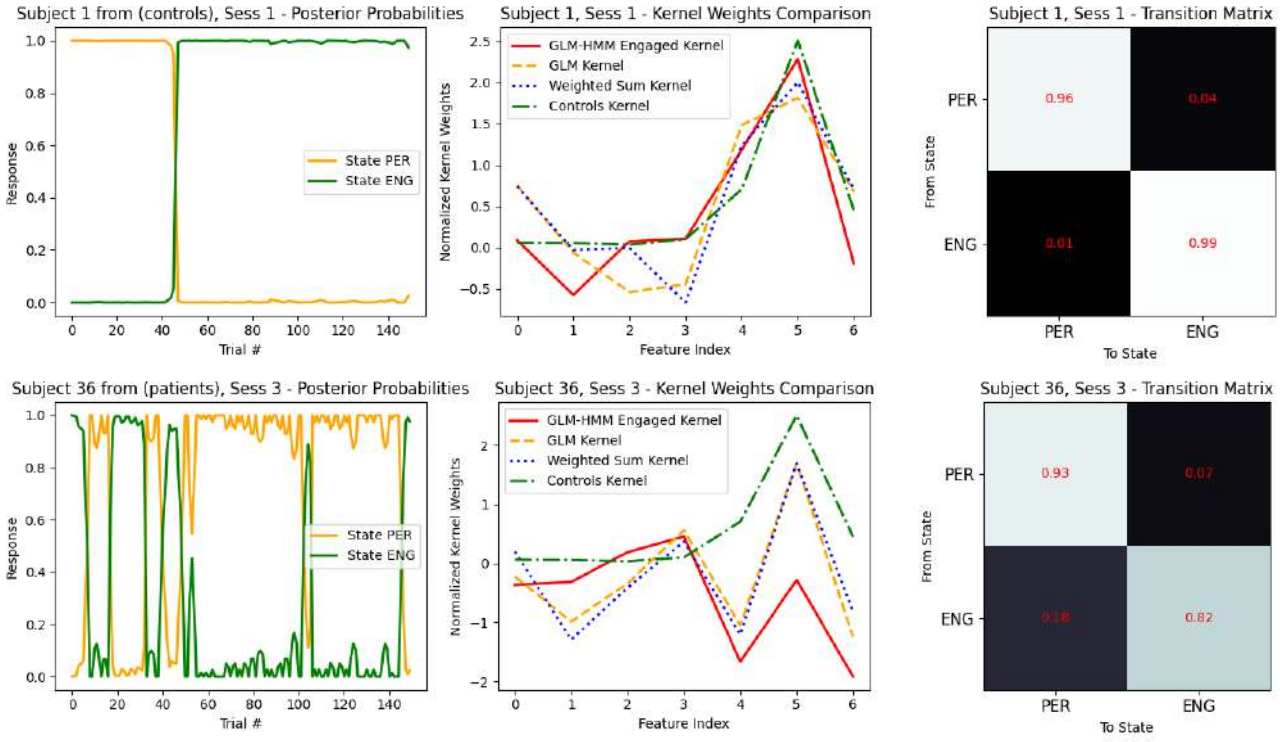


Fig. 8.1 **GLM-HMM state recovery and kernel estimation in a healthy subject (top row) and a stroke patient (bottom row) Left:** Posterior state probabilities across trials (n=150) for each participant. The control subject shows a stable engaged state (green) after a brief initial perseverative period (orange), whereas the patient alternates frequently between engaged and perseverative states throughout the session. **Center:** Comparison of normalized kernel weights for the engaged state estimated by GLM-HMM (red), the single-state GLM kernel (orange), the weighted sum kernel (blue), and the controls group average (green). The control's engaged-state kernel closely matches the group average, while the patient's kernel is more variable and deviates from typical controls. **Right:** State transition matrices reveal high state persistence for the control, but the patient exhibits reduced stability and increased switching between states, especially between perseverative and engaged periods.

ever, in patients, the correlation was significantly lower at 0.67, suggesting that including perseverative trials adds a moderate to strong impact on kernel estimation.

These findings highlight that while removing perseverative trials refines kernel estimation, it does not fully align the patient group with control-like kernel structures. This confirms that perseveration is not the only factor contributing to differences in kernel and suggests that additional impairments in stimulus processing and decision-making persist even when patients are classified as engaged.

Intra- and inter-individual analyses (Figures 8.25 and 8.26, at the end of this chapter)

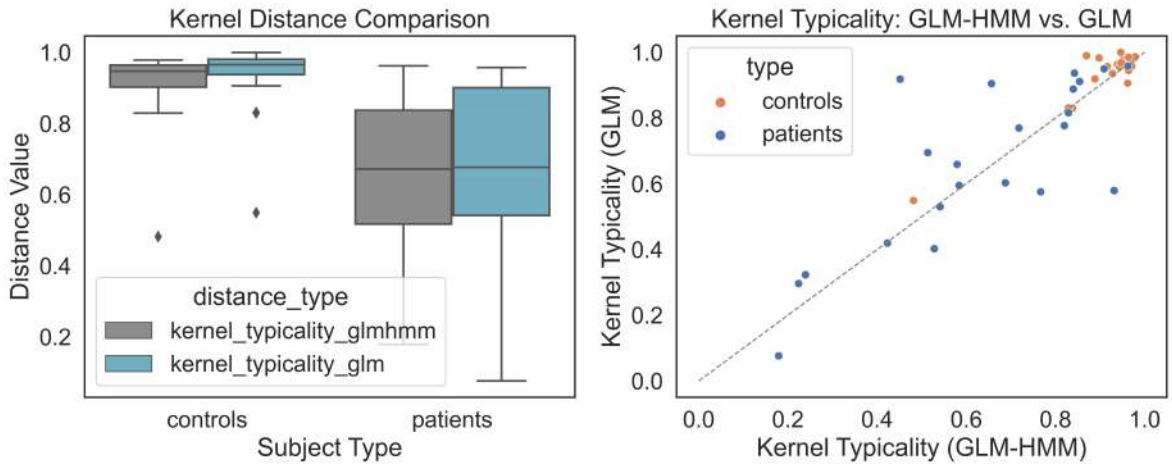


Fig. 8.2 **Kernel typicality analysis for controls and stroke patients Left:** Boxplots show the distribution of kernel distances (typicality) relative to the control group for each subject, as estimated by the GLM-HMM (engaged state only) and conventional GLM. Controls display consistently high kernel typicality, while patients show greater variability and generally lower values, indicating persistent deviations from normative stimulus-response mapping even after filtering out perseverative trials. **Right:** Scatter plot comparing kernel typicality scores from the GLM-HMM and GLM for each individual. Controls cluster near the identity line, confirming stable kernel estimates, whereas patients exhibit more scatter, further highlighting individual differences in perceptual precision and the limited effect of removing perseverative trials on restoring a stimulus-driven kernel.

further illustrate that kernel estimates for control participants remain relatively stable across the three estimation methods (CI, GLM, GLM-HMM), suggesting that perseveration minimally affects these individuals. In contrast, RH stroke patients display more variability in kernel shape and magnitude between methods, reflecting greater heterogeneity in response strategies and a stronger impact of perseveration on kernel estimation.

8.1.3 Impact on internal noise estimation

Beyond kernel estimation, we investigated internal noise as a potential biomarker, a concept thoroughly discussed in Chapter 6. Internal noise quantifies the variability in an observer's decision-making process that is not directly driven by the stimulus, making it a crucial measure of sensory uncertainty and cognitive stability.

Focusing on the engaged state, where decisions are primarily stimulus-driven, we

sought to determine whether internal noise levels remain stable or differ between patients and controls. If internal noise remains significantly elevated in patients, even when perseveration is removed, it would confirm that the high levels of noise seen in Chapter 4 are not overestimated because of perseveration but indicative of a broader impairment in sensory integration or cognitive control.

To assess this, we estimated internal noise before and after filtering perseverative trials, using the confidence interval of the GLM as a measure of variability (as proposed in Chapter 6).

Our findings (Figure 8.3) reveal a significant reduction in internal noise after removing perseverative trials, with 80% of cases (62 out of 78) showing lower internal noise values post-filtering. The average internal noise for controls decreases from 1.4 to 1.06, whereas for patients, it dropped substantially from 7.19 to 2.82. This suggests, first, that perseveration contributes significantly to increased estimated internal noise in this group, i.e., that much of the apparent decision variability in patients is not due to inherent noise but rather to lapses in attention characterized by perseverative responding. This distinction is important: it implies that interventions targeting sustained attention and strategies to reduce perseveration could meaningfully improve perceptual decision-making in stroke patients. Second, and nonetheless, even after excluding perseverative episodes, patients' internal noise remains higher than that of controls, suggesting that additional impairments in sensory integration or cognitive control persist.

8.2 Comparison between patients and controls on the new estimates

8.2.1 Kernel typicality

A key result of Chapter 4 was that kernel typicality strongly differentiated patients and controls. By identifying perseverative trials, the GLM-HMM method aims to obtain a more accurate estimation of the GLM kernel for engaged trials.

Figure 8.4 shows how old (CI, GLM) and new (GLM-HMM) methods differ in compar-

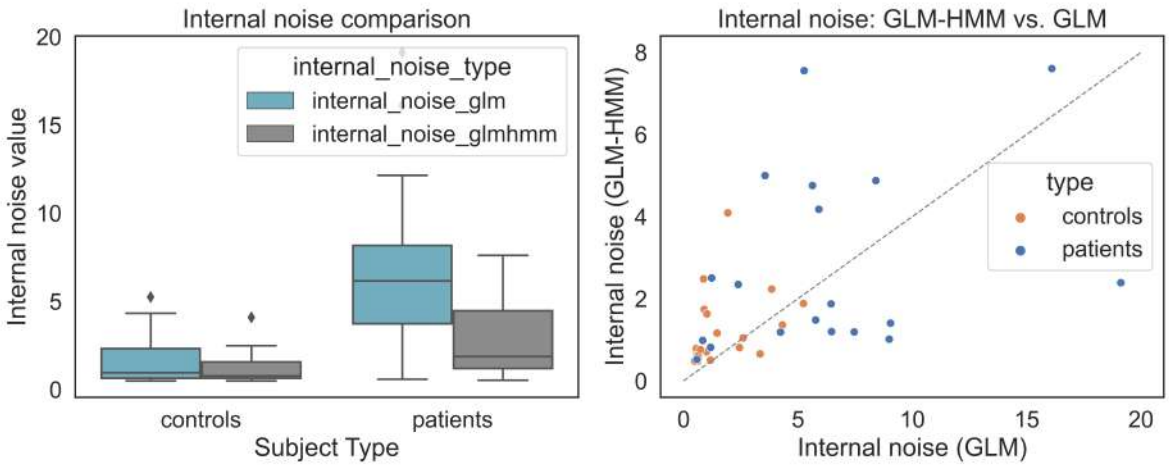


Fig. 8.3 **Comparison of internal noise estimates for controls and patients, before and after removing perseverative trials Left:** Boxplots show that patients exhibit higher internal noise than controls when measured with standard GLM, but internal noise is significantly reduced for both groups, especially patients, when estimated using only engaged-state trials with GLM-HMM. **Right:** Scatterplot comparing internal noise values from GLM-HMM (y-axis) and GLM (x-axis) for each participant. Most points fall below the diagonal, indicating that internal noise decreases after excluding perseverative trials. This reduction is especially pronounced in patients, suggesting that perseveration inflates estimates of sensory uncertainty and that filtering out these trials yields a more accurate measure of intrinsic perceptual noise.

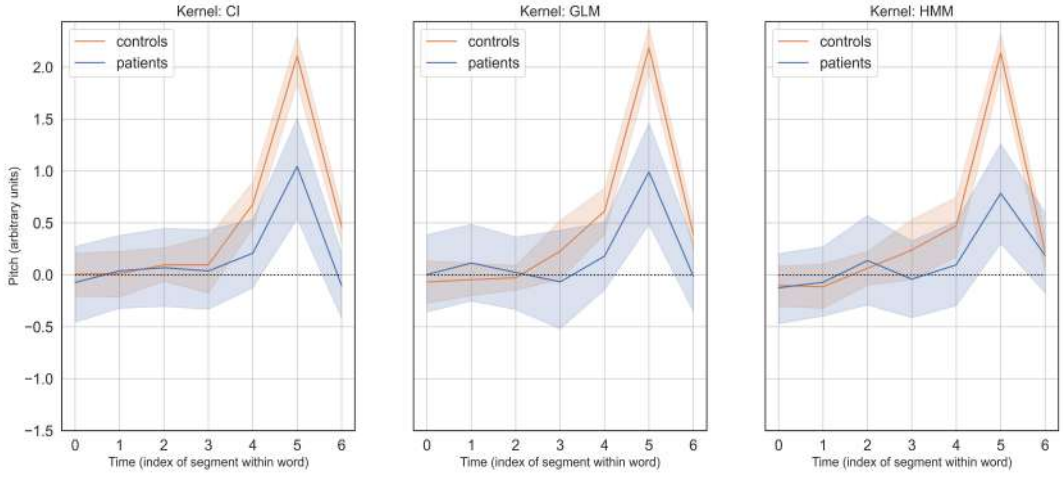


Fig. 8.4 **Comparison of internal representations (kernels) for controls and patients, estimated using three different methods** Classification Image (CI, left), GLM (middle), and GLM-HMM (right, based only on engaged trials). For all methods, kernels are plotted as a function of time (segment index within word).

ing patient and control kernels. The main difference is that after removing perseverative trials, estimated kernel magnitudes for patients tend to be lower, more uncertain, and more different from controls than with previous methods.

It appears unlikely that *including* perseverating trials in kernel computation accidentally made patient responses more similar to controls. A more likely interpretation is that a significant proportion of patient responses occur in the perseverative state. Specifically, while controls have an average of 119 engaged trials out of 150 (80%), patients have significantly fewer, with only 62 engaged trials out of 150 (41%). This substantial reduction in trial count affects kernel estimation, as seen, for instance, in Figure 3.3, limiting our ability to obtain robust kernel estimates.

In figure 8.5, we present a scatter plot comparing patients' kernel typicality in dark blue and how closely their kernel resembles that of a control group in orange across three different methods: kernel-CI (Classification Images or weighted sum) for all trials, kernel-GLM (Generalized Linear Model) for all trials, and kernel-HMM, which represents the GLM kernel computed only for engaged trials. The size of each point represents the number of engaged trials for each participant, binary classified as "less engaged" or "more engaged" based on the number of engaged trials (cut-off value: n = 114/150 for controls, n = 70/150 patients). Our findings indicate that patients with fewer engaged trials tend to deviate further from the diagonal line, which represents the similarity between methods. This effect is particularly evident in the third plot, where patients with fewer engaged trials show a pronounced divergence from the diagonal, highlighting the impact of engagement level on kernel similarity across methods.

8.2.2 Internal noise

Figure 8.6 shows scatter plots comparing internal noise across different estimation methods.

We observe distinct patterns between patients and controls. The Double-Pass (DP) method, which does not account for trial-by-trial state transitions, tends to overestimate internal noise in patients with low engagement, as seen in the upper-left quadrant of

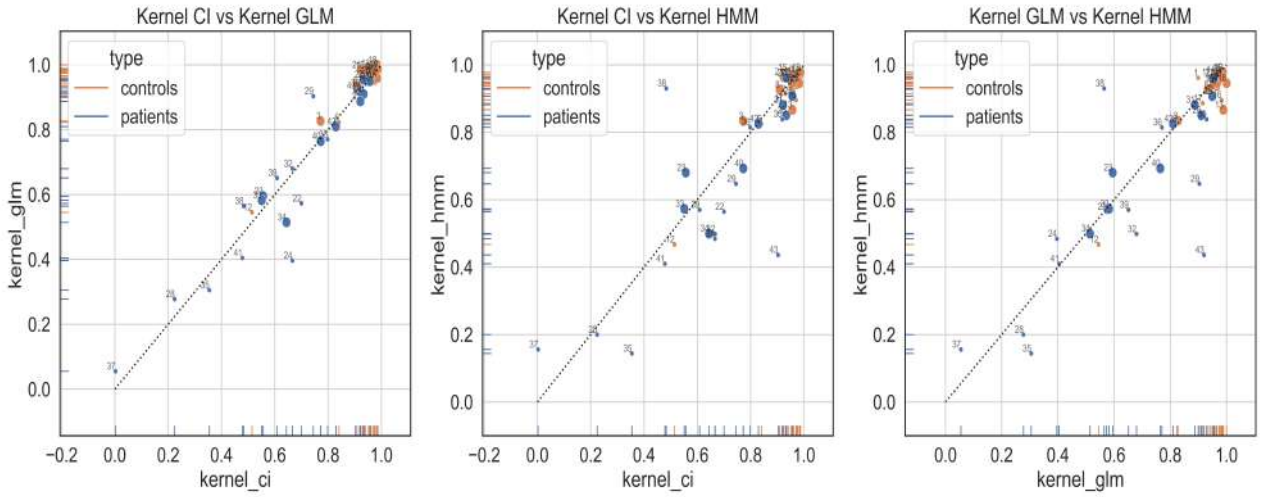


Fig. 8.5 **Scatter plots comparing kernel typicality between methods for controls (orange) and patients (blue) Left:** kernel-GLM vs. kernel-CI, **Middle:** kernel-HMM vs. kernel-CI, and **Right:** kernel-HMM vs. kernel-GLM. Each point represents a participant; the proximity to the diagonal indicates greater similarity in kernel typicality between methods. Controls generally cluster along the diagonal, reflecting high agreement across methods. In contrast, many patients, especially those with fewer engaged trials, deviate from the diagonal, most notably in the GLM-HMM vs. GLM comparison, illustrating that lower engagement leads to reduced method agreement and greater variability in kernel estimation.

the DP vs. GLM-HMM plot (figure 8.6). This suggests that many of these patients had perseverative trials that artificially inflated their noise estimates. In contrast, when using the GLM-HMM method, which isolates engaged trials, internal noise estimates decrease, particularly for low-engagement patients. Highly engaged patients, represented by larger points, exhibit more stable noise estimates across all methods, indicating that their decision-making is less affected by perseveration.

Similarly, the GLM method, which estimates noise across all trials without distinguishing between states, generally reports higher internal noise than GLM-HMM, particularly in patients with low engagement. This highlights the importance of accounting for perseverative trials in noise estimation, as their inclusion in GLM leads to inflated confidence intervals and higher uncertainty. Controls, on the other hand, consistently show low internal noise across all methods, aligning closely with the diagonal in all plots. The absence of points in the lower-right quadrant of the DP vs. GLM-HMM plot confirms that patients with low DP noise also have low noise when assessed through engaged trials, reinforcing

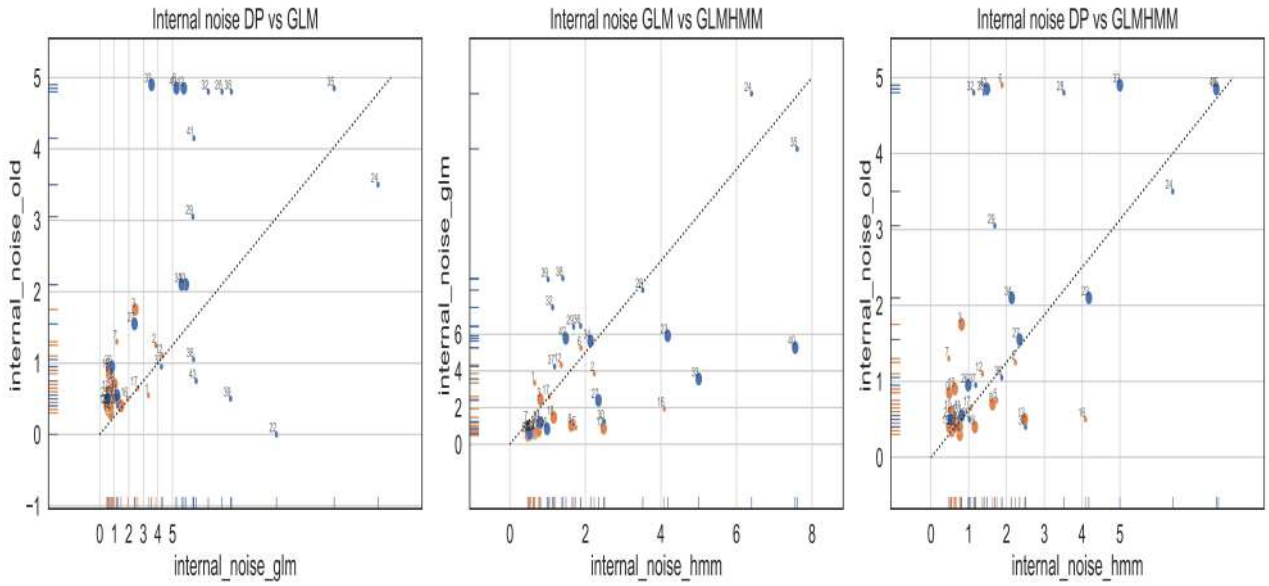


Fig. 8.6 **Scatter plots comparing internal noise estimates across three methods: Double-Pass (DP) vs. GLM (left), GLM vs. GLM-HMM (middle), and DP vs. GLM-HMM (right), for controls (orange) and patients (blue).** Each point represents a participant, with point size indicating the number of engaged trials. In patients, both DP and GLM methods, when applied across all trials, tend to overestimate internal noise, especially for those with lower engagement, as shown by points deviating above the diagonal. The GLM-HMM method, which isolates engaged trials, yields lower and more accurate internal noise estimates for these patients, bringing their values closer to those of controls. Highly engaged participants (larger points) show consistent noise estimates across all methods. These results highlight the importance of accounting for trial-by-trial fluctuations in engagement, as failing to do so leads to inflated internal noise estimates in patient populations due to perseverative responding.

the idea that noise, when present, is pervasive rather than trial-specific.

Overall, these findings confirm that perseveration significantly impacts internal noise estimation in patients, leading to overestimation in DP and GLM methods. The GLM-HMM approach provides a more refined measure by focusing only on engaged trials, revealing that some patients had artificially high noise estimates due to perseverative responses. This highlights the importance of separating engaged from perseverative trials to obtain a more precise assessment of response consistency and decision-making reliability in patient populations.

8.2.3 A new biomarker: transition probabilities

Our analysis of transition matrices for both a healthy participant and a stroke patient reveals distinct differences in state-switching behavior between patients and controls. One key question we explored was whether patients exhibit more frequent state transitions than controls, and if so, in which direction.

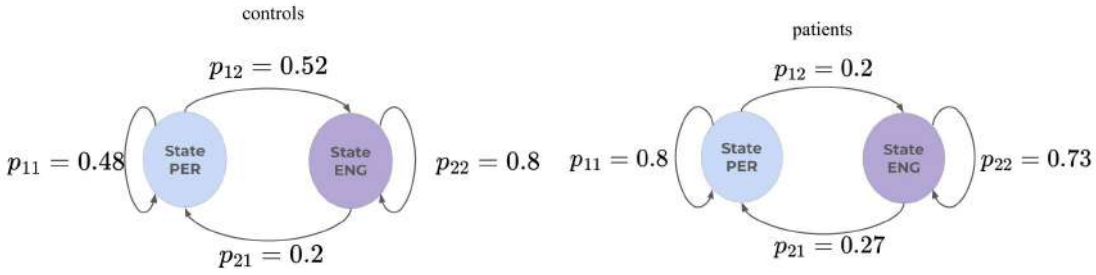


Fig. 8.7 **Schematic illustration of average state transition probabilities for controls (left) and patients (right).** Arrows indicate the probability of remaining in or transitioning between the engaged (ENG) and perseverative (PER) states. Controls are more likely to remain in the engaged state ($p_{22} = 0.8$), while patients tend to persist in the perseverative state ($p_{11} = 0.8$) and are less likely to transition back to engagement ($p_{21} = 0.27$ for patients vs. $p_{21} = 0.2$ for controls). These results highlight group differences in the dynamics of state switching.

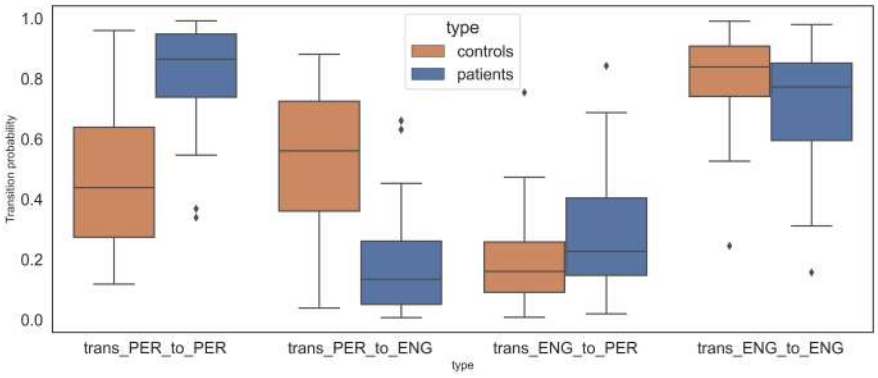


Fig. 8.8 **Comparison of state transition probabilities for controls (orange) and patients (blue).** Patients exhibit significantly higher probabilities of remaining in the perseverative state (PERtoPER) and lower probabilities of switching from perseverative to engaged state (PERtoENG) compared to controls. No significant group difference is observed for transitions from engaged to perseverative (ENGtoPER) or engaged to engaged (ENGtoENG) states.

On average, patients show a slightly higher tendency to perseverate, remaining in the perseverative state with a probability of 0.27 (equivalently, transitioning back to the

engaged state with a probability of 0.73). In contrast, controls exhibit a lower, but not inexistent, probability of perseveration (0.2) or, equivalently, a higher probability of maintaining the engaged state (0.8).

The main difference, however, emerges when both groups enter the perseverative state. Controls tend to exit perseveration more easily, with a 52% probability of returning to the engaged state (equivalently, a 48% probability of staying in perseveration). In contrast, patients show a strong persistence in perseveration, remaining in this state with a probability of 0.8 or, equivalently, transitioning back to the engaged state only 20% of the time.

This suggests that the difference between patient and control is not their tendency to decouple from the stimulus, which is only moderately higher in patients, but rather that, once patients enter a perseverative mode, they struggle to re-engage with the stimulus, whereas controls demonstrate greater flexibility in state switching (Figure 8.8).

An alternative view on the same probabilities concerns dwell time, which represents the number of consecutive trials spent in each state. In Markov processes, dwell time d in a state i obeys an exponential probability distribution, $p_i(d)$ which is a direct function of state self-transition probability a_{ii}:

$$p_i(d) = a_{ii}^{d-1}(1 - a_{ii}) \tag{8.1}$$

where $\overset{p}{i}(d)$ is the probability of d consecutive observations in a state i (Rabiner & Juang, 1986).

On average (Figure 8.9), healthy participants remain in the engaged state for 10 trials, whereas stroke patients stay engaged for only 7 trials. In contrast, controls tend to exit the perseverative state quickly, staying in it for an average of just 2 trials, while patients remain in the perseverative state for a significantly longer duration, averaging 10 trials. This confirms the probability observations above: patients enter the perseverative state moderately more frequently but strikingly struggle to disengage from it once they are in it.

In addition, we investigated whether transition probabilities are related to old biomarkers kernel typicality and internal noise by performing a regression analysis on data from

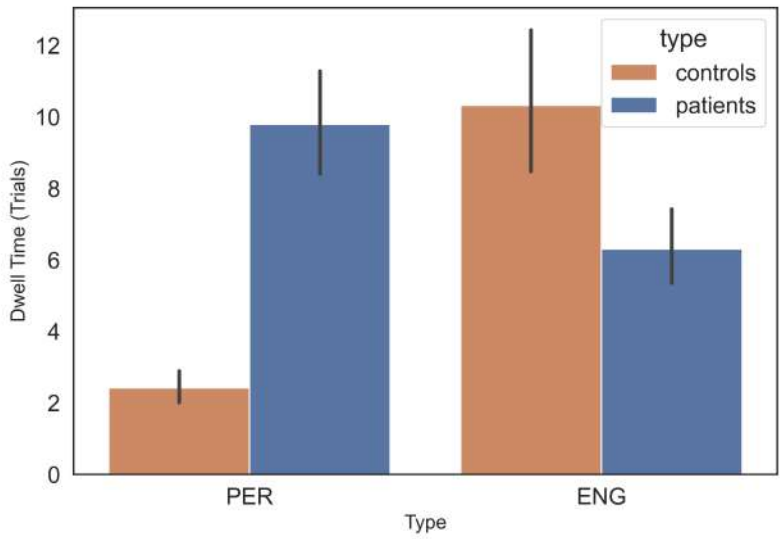


Fig. 8.9 **Average dwell time (in trials) spent in the perseverative (PER) and engaged (ENG) states for controls (orange) and patients (blue).** Controls typically remain in the engaged state for around 10 trials and exit the perseverative state quickly (average of 2 trials). In contrast, stroke patients show a much longer dwell time in the perseverative state (an average of 10 trials) and a shorter dwell time in the engaged state (7 trials), indicating both greater persistence in perseveration and greater difficulty maintaining engagement.

both patients and controls. The goal was to determine whether transition probabilities could predict kernel typicality and internal noise levels. Our results indicate that the probability of transitioning to the engaged state is significantly associated with kernel typicality across all three kernel estimation methods (CI, GLM, and HMM), as well as with internal noise estimated by GLM. However, no significant relationships were found between transition probability to the perseverative state and kernel typicality or internal noise. This aligns with our previous findings, where the probability of entering perseveration did not significantly differ between patients and controls, suggesting that the key distinguishing factor is not how often individuals enter perseveration but rather how effectively they can transition back to an engaged state.

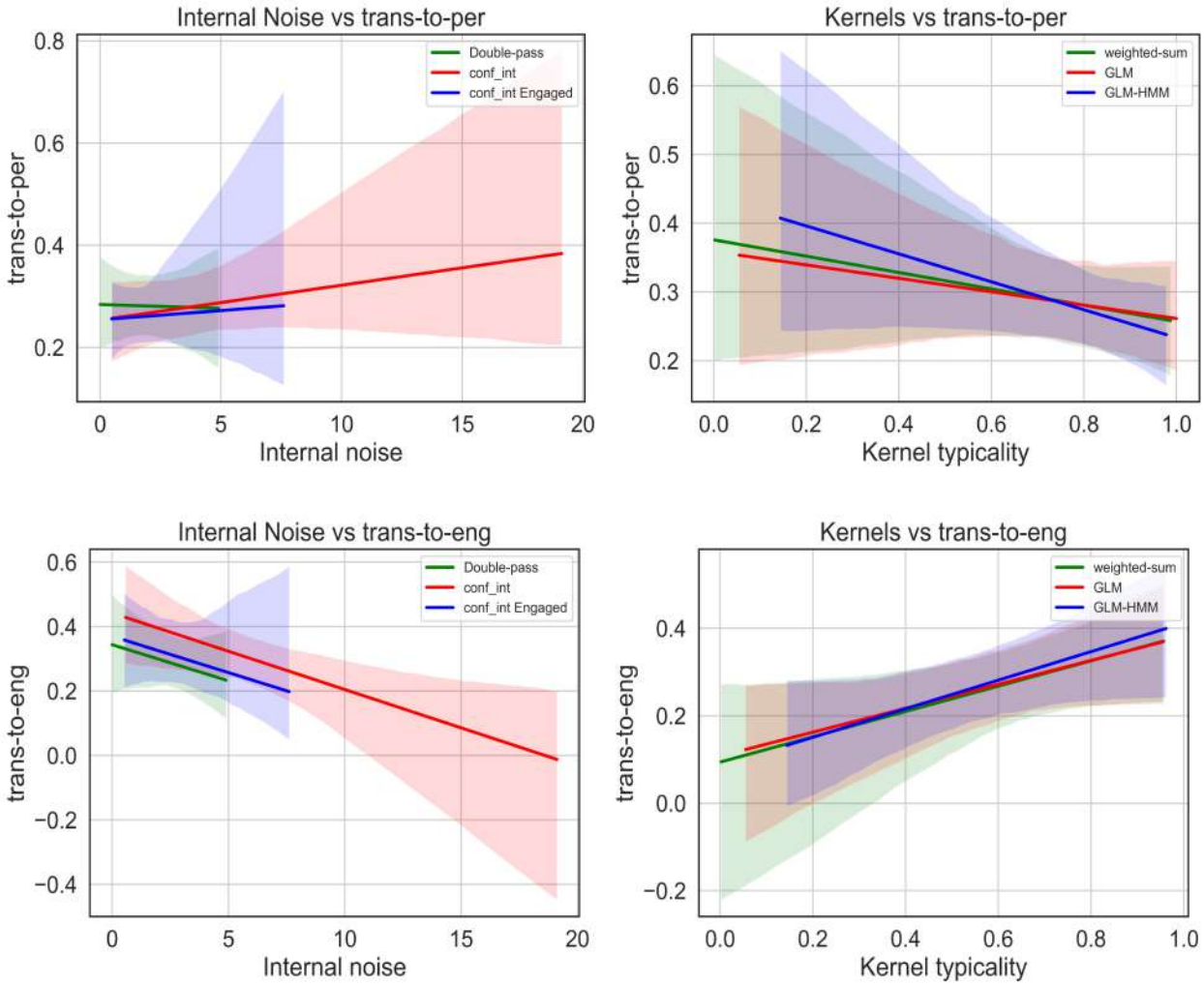


Fig. 8.10 **Relationship between transition probabilities, internal noise, and kernel typicality. Top row:** Probability of transitioning to the perseverative state ("trans-to-per") as a function of internal noise (left) and kernel typicality (right). **Bottom row:** Probability of transitioning to the engaged state ("trans-to-eng") as a function of internal noise (left) and kernel typicality (right). Results show that the probability of transitioning back to the engaged state is positively associated with kernel typicality and negatively associated with internal noise, particularly for model-based estimates. In contrast, transition probability to the perseverative state shows no clear association with either kernel typicality or internal noise.

8.3 Comparison between symptomatic and asymptomatic patients

Another key result of Chapter 4 is that kernel typicality and internal noise allowed to differentiate controls from patients evaluated at MEC (Montreal Evaluation of Communication) > 9, which is the cut-off for pathological aprosody. We examine here how removing perseverating trials affects that result.

8.3.1 Kernel typicality

In Figure 8.11, we present patients classified based on their pathological score in the MEC. As shown in the scatter plot, we defined patients according to a cut-off score of 9, with those scoring above this threshold (MEC > 9) represented in dark blue, indicating no pathological impairment, while those scoring 9 or below (MEC≤ 9) are shown in light blue, indicating a prosody perception deficit.

Similar to Figure 8.5, the size of the points in the plot represents the number of engaged trials per patient. On average, patients with MEC > 9 had 71 engaged trials, whereas patients with MEC≤ 9 had an average of 69 engaged trials. Regarding kernel typicality across methods, control participants showed consistently high values across all approaches, with kernel-CI at 0.92, kernel-GLM at 0.93, and kernel-HMM at 0.91. Among patients with MEC≤ 9, kernel typicality was lower, with values of 0.61 for kernel-CI, 0.57 for kernel-GLM, and 0.57 for kernel-HMM. Patients with MEC > 9 exhibited intermediate kernel typicality, with scores of 0.74 for kernel-CI, 0.75 for kernel-GLM, and 0.68 for kernel-HMM. These findings suggest that kernel typicality decreases in patients with greater prosody perception deficits, particularly when estimated using the GLM or HMM methods.

Mann-Whitney U tests further confirm these differences. Figure 8.12 shows significant distinctions between patients and controls but weaker effects when comparing MEC≤ 9 and MEC > 9 patients.

Comparing MEC≤ 9 vs. MEC > 9 patients, there is no significant difference in kernel

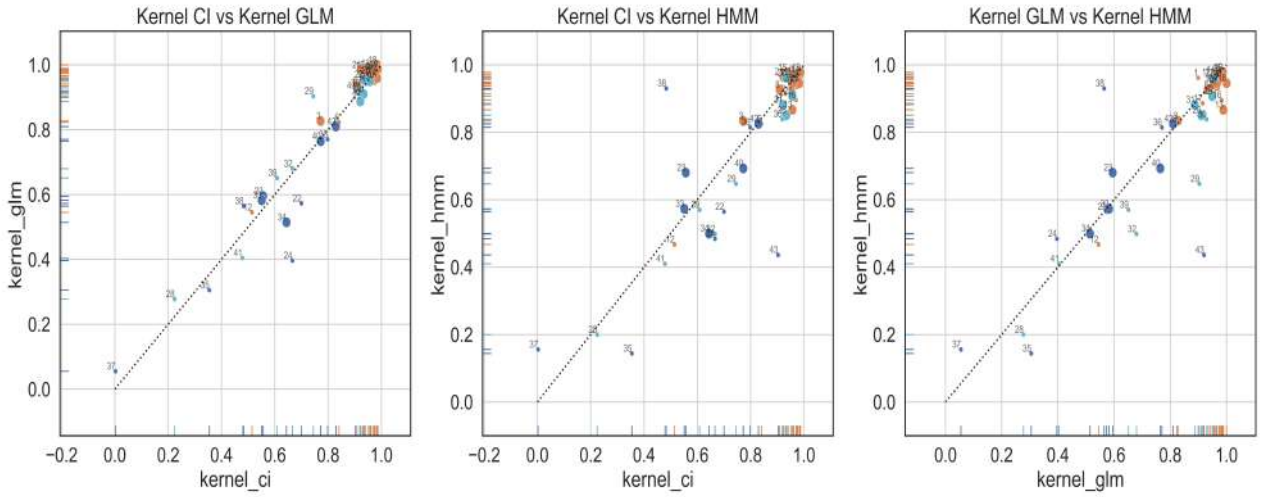


Fig. 8.11 **Scatter plots comparing kernel typicality between methods for controls (orange), patients with MEC > 9 (dark blue), and patients with MEC ≤ 9 (light blue).Left:** kernel-GLM vs. kernel-CI. **Center:** kernel-HMM vs. kernel-CI.**Right:** kernel-HMM vs. kernel-GLM. The size of each point reflects the number of engaged trials per participant. Controls consistently show high kernel typicality and strong agreement between methods, clustering near the diagonal. In contrast, patients, especially those with greater prosodic deficits (MEC ≤ 9, light blue), exhibit reduced kernel typicality, increased variability, and more points deviating from the diagonal. This pattern highlights that prosody perception deficits are associated with reduced kernel stability and consistency across estimation approaches.

typicality across all methods ($p > 0.05$). This result should be interpreted with caution, as the MEC may lack the sensitivity to detect subtle prosodic deficits; thus, even patients with MEC > 9 may still present with underlying impairments that go unrecognized by MEC.

Comparing MEC≤ 9 patients to controls, we find highly significant differences across all kernel estimation methods ($p = 0.000$), with patients showing significantly lower kernel typicality. This indicates that patients with severe prosody perception deficits exhibit markedly different kernel structures compared to controls.

Comparing MEC > 9 patients to controls, there are significant but less pronounced differences ($p = 0.002$ - 0.004), indicating that patients categorized by MEC having mild or no prosody deficits still show reduced kernel typicality compared to controls, but to a lesser extent than MEC≤ 9 patients. There are no differences in the three estimation methods at the level of kernel typicality.

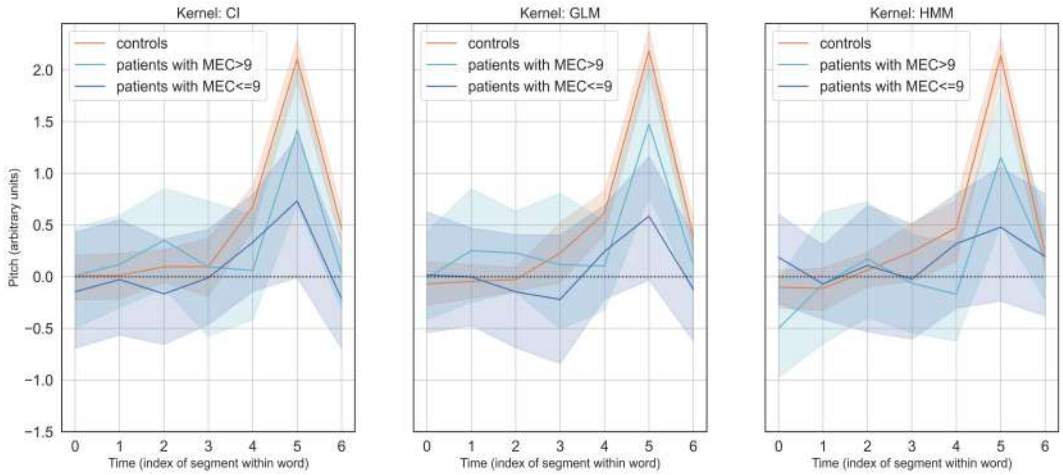


Fig. 8.12 **Comparison of internal representations (kernels) across groups and estimation methods.** For each method, Classification Image (CI, left), GLM (middle), and GLM-HMM (right, based only on engaged trials), kernels are plotted as a function of time (segment index within word). Results are shown separately for controls (orange), patients categorized by MEC with mild or no prosodic deficit (MEC > 9, dark blue), and patients categorized by MEC with severe prosodic impairment (MEC ≤ 9, light blue). Kernels are plotted as a function of time (segment index within word). For both patient groups, kernel shapes remain similarly atypical relative to controls, regardless of whether perseverative trials are removed (GLM-HMM), suggesting that filtering out these trials does not markedly improve the quality of internal representations in patients with prosodic deficits.

Finally, a similar analysis can be done by separating the comprehension and repetition subtasks of MEC. For patients with normal (or near-normal) prosodic performance on MEC comprehension task(MEC-c > 9), kernel typicality was significantly lower than controls across all methods: Kernel CI ($p = 0.001$, $r = -0.67$), Kernel GLM ($p = 0.001$, $r = -0.70$), and Kernel HMM ($p = 0.001$, $r = -0.73$). For patients with normal (or near-normal) prosodic performance on MEC repetition task (ME-r > 9), kernel typicality was significantly lower across all methods: Kernel CI ($p = 0.000$, $r = -0.80$), Kernel GLM ($p = 0.000$, $r = -0.83$), and Kernel HMM ($p = 0.000$, $r = -0.84$).

8.3.2 Internal noise

The internal noise estimates across different methods reveal key distinctions between control participants and patients, particularly those with prosody perception deficits

(MEC$\leq$ 9). In the Double-Pass (DP) method, there is a significant difference between patients with MEC$\leq$ 9 and controls ($p = 0.004$), with patients exhibiting much higher noise levels (2.89 vs. 0.89). However, no significant difference is found between MEC$\leq$ 9 and MEC $>$ 9 patients ($p = 0.409$), suggesting that DP does not capture within-patient variability effectively.

This difference persists in the GLM-based estimation, but with even greater separation, as noise in the MEC$\leq$ 9 group rises to 8.33, compared to 4.4 in MEC $>$ 9 patients and 1.63 in controls. The GLM method shows a highly significant difference between MEC$\leq$ 9 and controls ($p = 0.000$), as well as between MEC $>$ 9 and controls ($p = 0.027$), indicating that patients, regardless of severity, exhibit higher noise levels than controls. However, no significant difference is found between MEC$\leq$ 9 and MEC $>$ 9 patients ($p = 0.138$), reinforcing that GLM captures overall group-level differences but not fine-grained variations within patients. The fact that GLM estimates higher noise suggests that perseveration inflates confidence intervals when all trials are considered.

When using the GLM-HMM method, which isolates engaged trials, internal noise in MEC$\leq$ 9 patients is significantly reduced to 3.88, suggesting a more accurate measure of perceptual variability. Despite this reduction of internal noise, and unlike DP and GLM, GLM-HMM not only maintains a significant difference between MEC$\leq$ 9 and controls ($p = 0.001$) but also between MEC$\leq$ 9 and MEC $>$ 9 patients ($p = 0.030$). This may be because perseveration in MEC $>$ 9 patients was more prevalent in repeated trials, leading to inflated noise estimates in DP and GLM. However, no significant difference is found between MEC $>$ 9 and controls in GLM-HMM ($p = 0.133$), indicating that after removing perseverative trials, the noise levels of MEC $>$ 9 patients resemble those of controls.

As above, a similar analysis can be done by separating the comprehension and repetition subtasks of MEC. For patients with normal (or near-normal) prosodic performance on MEC comprehension task(MEC-c $>$ 9), there was a significant difference in the GLM internal noise estimate ($p = 0.006$, $r = 0.58$), where patients showed higher internal noise than controls. However, other internal noise estimates (DP, old method, and HMM) showed only marginal trends ($p = 0.053$ - 0.064), without reaching significance. For patients with normal (or near-normal) prosodic performance on MEC repetition task (ME-r

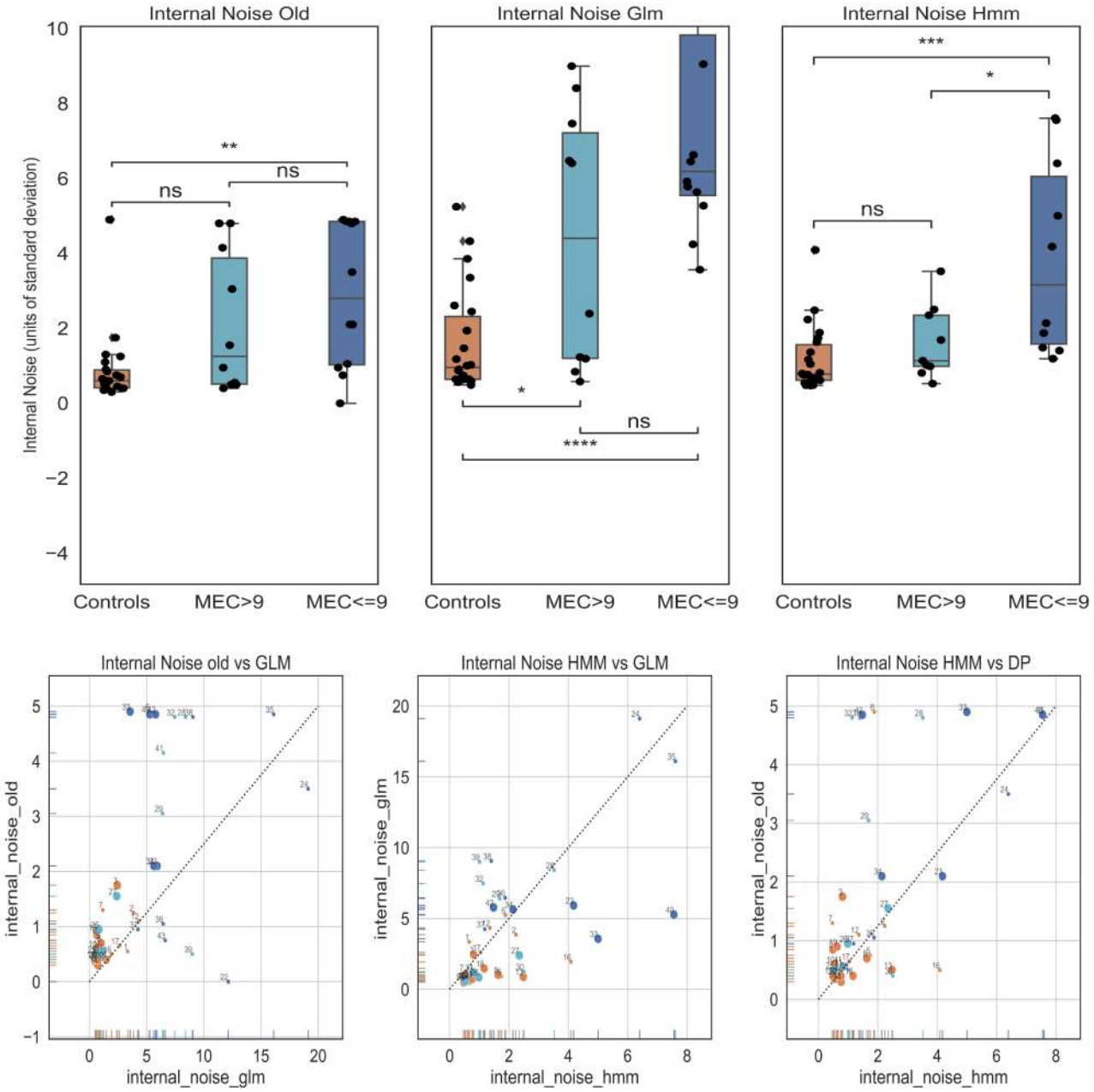


Fig. 8.13 **Internal noise estimates across three methods, Double-Pass (Old), GLM, and GLM-HMM for controls, patients with MEC > 9, and patients with MEC ≤ 9.** Patients with MEC ≤ 9 (greater prosody perception deficits) show significantly higher internal noise than controls in all methods, but the difference between patient subgroups is most pronounced in GLM-HMM, which isolates engaged trials. GLM-HMM reveals a significant reduction in internal noise for MEC > 9 patients compared to MEC ≤ 9 patients, and brings their noise levels closer to those of controls, suggesting that much of the excess noise in standard GLM and DP estimates stems from perseverative trials.

> 9), internal noise was significantly higher with GLM ($p = 0.000$, $r = 0.76$), HMM ($p = 0.007$, $r = 0.52$), and DP ($p = 0.015$, $r = 0.45$), all showing increased noise levels compared to controls.

8.3.3 Transition probabilities

Figure 8.14 shows transition probabilities for ENG and PER states across controls and patient groups.

Comparing MEC≤ 9 and MEC > 9 patients, no significant difference was found in the probability of transitioning to a perseverative state ($p = 0.079$) or to a non-perseverative state ($p = 0.206$), suggesting that both patient groups exhibit similar patterns of state switching, regardless of the severity of their prosody deficits.

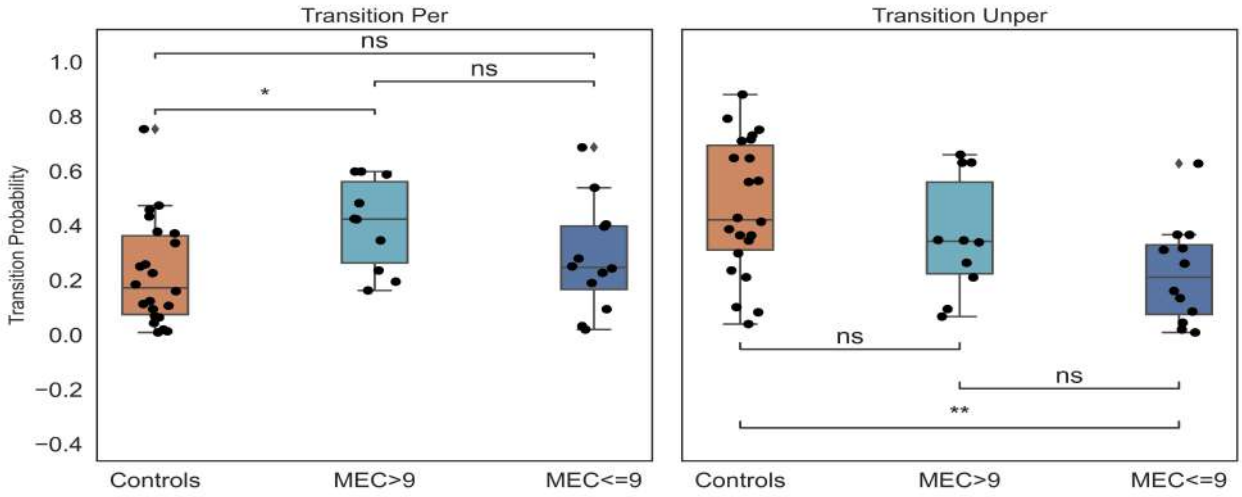


Fig. 8.14 **Transition probabilities for perseverative (left) and non-perseverative (right) states across controls and patient groups (MEC > 9 and MEC ≤ 9). Left:** The probability of entering the perseverative state ("Transition Per"), **Right:** The probability of transitioning to the engaged (non-perseverative) state ("Transition Unper"). Controls exhibit significantly lower perseverative transition probabilities compared to MEC > 9 patients, but there is no significant difference between the two patient groups. In contrast, MEC ≤ 9 patients display a significantly lower probability of non-perseverative transitions than both controls and MEC > 9 patients, indicating greater difficulty in returning to or sustaining engagement.

When comparing MEC≤ 9 patients to controls, no significant difference was observed in the probability of transitioning into a perseverative state ($p = 0.792$), indicating that

these patients are not necessarily more prone to perseveration than controls. However, their non-perseverative transition probability was significantly lower ($p = 0.022$), meaning that once in a perseverative state, MEC≤ 9 patients have more difficulty switching out of it and returning to an engaged state. This pattern points to a deficit in cognitive flexibility within this group.

For MEC > 9 patients compared to controls, a significant difference was found in perseverative transition probabilities ($p = 0.025$), with MEC > 9 patients showing fewer transitions into a perseverative state. This suggests that these patients, who have less severe or no prosody perception deficits (MEC > 9), are more resistant to entering perseveration, resembling control participants. Additionally, no significant difference was observed in their non-perseverative transition probabilities ($p = 0.249$), indicating that their ability to exit a perseverative state is similar to that of controls.

Taken together, these findings highlight that MEC≤ 9 patients exhibit greater difficulty disengaging from perseverative states, whereas MEC > 9 patients transition less frequently into perseveration, aligning more closely with controls. This suggests that transition probabilities could serve as a biomarker for perseveration severity. These transition probabilities likely reflect broader aspects of cognitive flexibility or executive functioning, beyond decision-making alone, with reduced non-perseverative transitions suggesting greater rigidity in MEC≤ 9 patients and fewer perseverative transitions indicating greater cognitive flexibility in MEC > 9 patients.

As above, a similar analysis can be done by separating the comprehension and repetition subtasks of MEC. For patients with normal (or near-normal) prosodic performance on MEC comprehension task(MEC-c > 9), no significant differences were found in transition probabilities for perseverative or non-perseverative states ($p > 0.05$). For patients with normal (or near-normal) prosodic performance on MEC repetition task (ME-r > 9), transitions to non-perseveration were significantly lower in this group ($p = 0.010$, $r = -0.50$), suggesting greater difficulty in returning to an engaged state. However, the difference in perseverative transition probability did not reach significance ($p = 0.069$).

8.4 Correlations with clinical measures within the patient group

In this section, we re-evaluate the various correlations found between reverse correlation parameters (within the patient group) and clinical measures using the new GLM-HMM estimates, as well as investigate whether transition probabilities are also sensitive to the same or other clinical factors. A summary of all statistics reported in this section can be found in Tables 8.1,8.2 and 8.3.

8.4.1 Correlation with MEC

We analyzed the relationship between GLM-HMM-derived measures and MEC scores to determine whether kernel typicality, internal noise, or transition probabilities were predictive of performance on clinical prosody subtests by MEC.

The regression results (Figure 8.15) indicate no significant correlation between MEC scores and kernel typicality (CI, GLM, or HMM), internal noise, or transition probabilities ($p > 0.05$), suggesting that these metrics do not strongly predict the MEC performance.

For kernel typicality, the relationships were positive but non-significant, meaning that higher kernel typicality does not systematically relate to better MEC scores (Kernel CI: $\beta = 2.93$, $p = 0.40$; Kernel GLM: $\beta = 3.34$, $p = 0.33$; Kernel HMM: $\beta = -1.77$, $p = 0.61$). For internal noise, the relationships were negative, indicating that higher internal noise may be associated with lower MEC scores, though not significantly (double-pass method: $\beta = -0.80$, $p = 0.07$; GLM: $\beta = -0.26$, $p = 0.16$; HMM: $\beta = -0.23$, $p = 0.60$). Notably, the old internal noise measure was nearly significant but is no longer after removing perseverative trials in HMM, suggesting that previously observed effects may have been inflated by perseveration rather than reflecting 'true' perceptual variability.

For transition probabilities, higher transition to the engaged state ($\beta = 5.97$, $p = 0.18$) was positively associated with MEC scores, suggesting that more frequent return to an engaged state may weakly relate to better communication ability, though not significantly. Transition to perseveration was also positively associated ($\beta = 1.04$, $p = 0.81$).

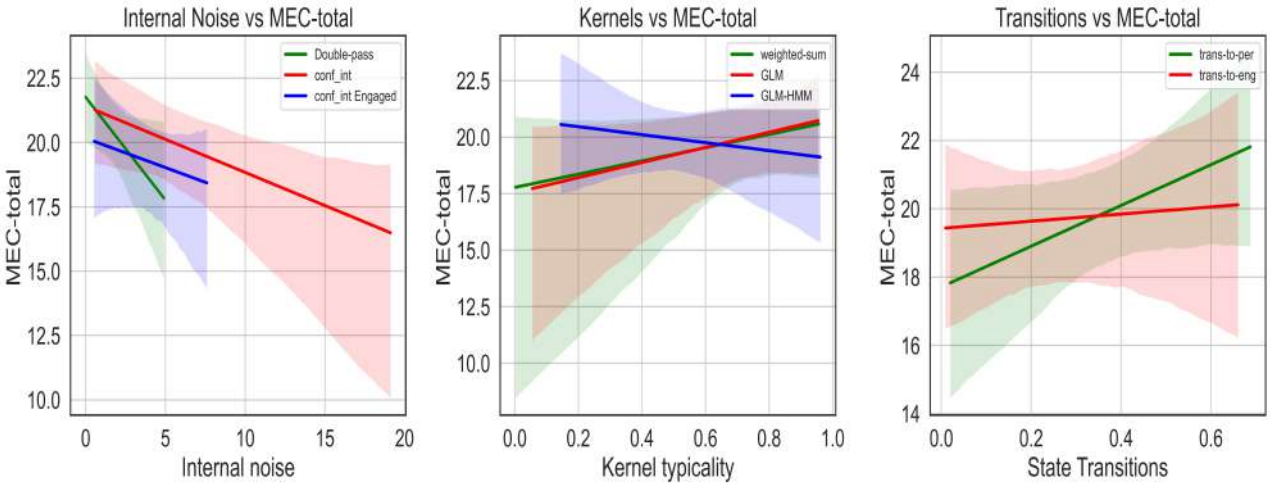


Fig. 8.15 **Relationship between behavioral metrics and clinical prosody performance (MEC-total).** MEC-total (Y-axis, all panels) sums two subtests (comprehension and repetition, 12 points each; max = 24). **(Left)** Internal noise: Double-pass (DP, old) estimates (0–5) and model-based GLM/GLM-HMM (new) estimates (up to 20, SD units); show negative, non-significant association with MEC-total. **(Middle)** Kernel typicality (0–1) : both weighted-sum (old) and GLM-HMM (new, engaged trials only) show positive, non-significant relationships with MEC-total. **(Right)** Transition probabilities (0–1, new): probability of transitioning to engaged and perseverative states show positive, non-significant relationships with MEC-total.

When focusing on prosody comprehension (MEC-C, Figure 8.16), we found a significant negative correlation between MEC-C and internal noise from the double-pass method ($\beta = -0.62$, $p = 0.043$), which we in Chapter 4 took to suggest that patients with higher internal noise in double-pass trials may struggle more with prosody comprehension. However, internal noise from HMM was not significantly correlated, implying that this association in the old method might be driven by perseveration rather than true perceptual variability. Kernel typicality (Kernel CI: $\beta = 3.25$, $p = 0.18$; Kernel GLM: $\beta = 3.29$, $p = 0.16$; Kernel HMM: $\beta = -0.19$, $p = 0.94$) and transition probabilities (Perseverative: $\beta = 3.83$, $p = 0.22$; Engaged: $\beta = 3.27$, $p = 0.27$) showed no significant relationships with prosody comprehension.

For prosody repetition (MEC-R, Figure 8.17), none of the predictors showed a significant relationship with kernel typicality, internal noise, or state transitions ($p > 0.05$), indicating that repetition performance is not strongly associated with these computational measures. Notably, the internal noise old method previously showed a weak negative

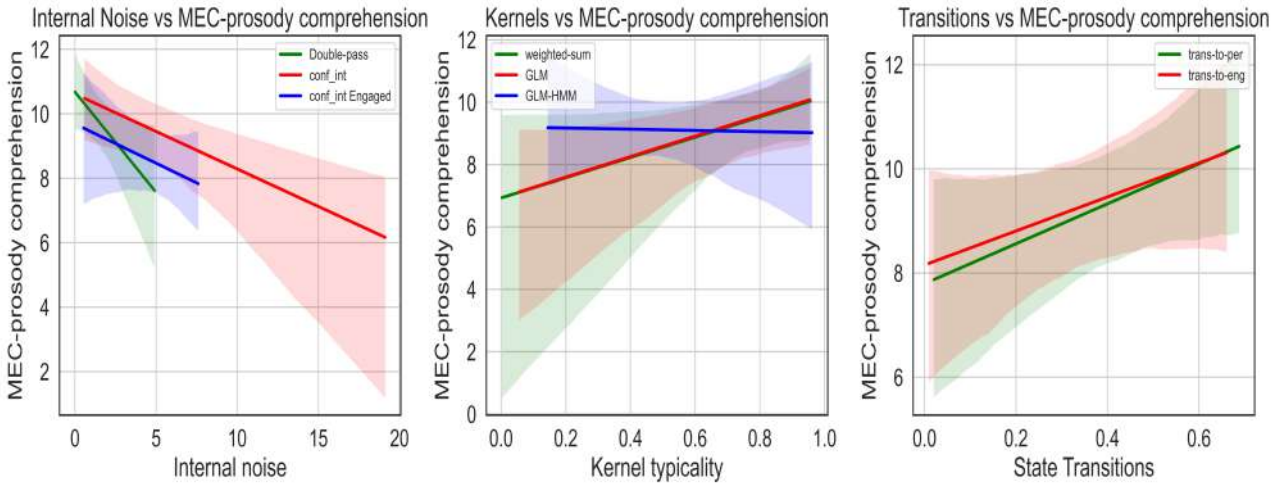


Fig. 8.16 **Relationship between behavioral metrics and prosody comprehension task (MEC-comprehension).** MEC-comprehension (Y-axis, all panels) is scored out of 12. **(Left)** Internal noise: Double-pass (DP, old) estimates show a significant negative association with MEC-comprehension; GLM and GLM-HMM model-based (new) estimates show no significant association. **(Middle)** Kernel typicality (0–1): Both weighted-sum (old) and GLM-HMM (new, engaged trials only) show positive, non-significant relationships with MEC-comprehension. **(Right)** Transition probabilities (0–1, new): probability of transitioning to engaged or perseverative states shows positive, non-significant associations with MEC-comprehension.

correlation ($\beta = -0.18$, $p = 0.37$), but this relationship is no longer present in HMM ($\beta = 0.02$, $p = 0.93$), suggesting once again that the previous measure captured perseveration effects rather than genuine perceptual noise.

8.4.2 Correlation with AIRTAC

Regression analysis with AIRTAC (central auditory processing ability) revealed strong significant relationships between kernel typicality and AIRTAC scores ($p < 0.01$, Figure 8.18), across methods. This indicates that higher kernel typicality is associated with better auditory processing (Kernel CI: $\beta = 12.08$, $p = 0.009$; Kernel GLM: $\beta = 11.28$, $p = 0.007$; Kernel HMM: $\beta = 12.72$, $p = 0.003$).

For internal noise, the old method was significantly negatively correlated with AIRTAC ($\beta = -1.20$, $p = 0.037$), meaning that higher noise in double-pass trials was associated with worse auditory discrimination. However, internal noise in HMM ($\beta = -0.46$, $p = 0.49$) and

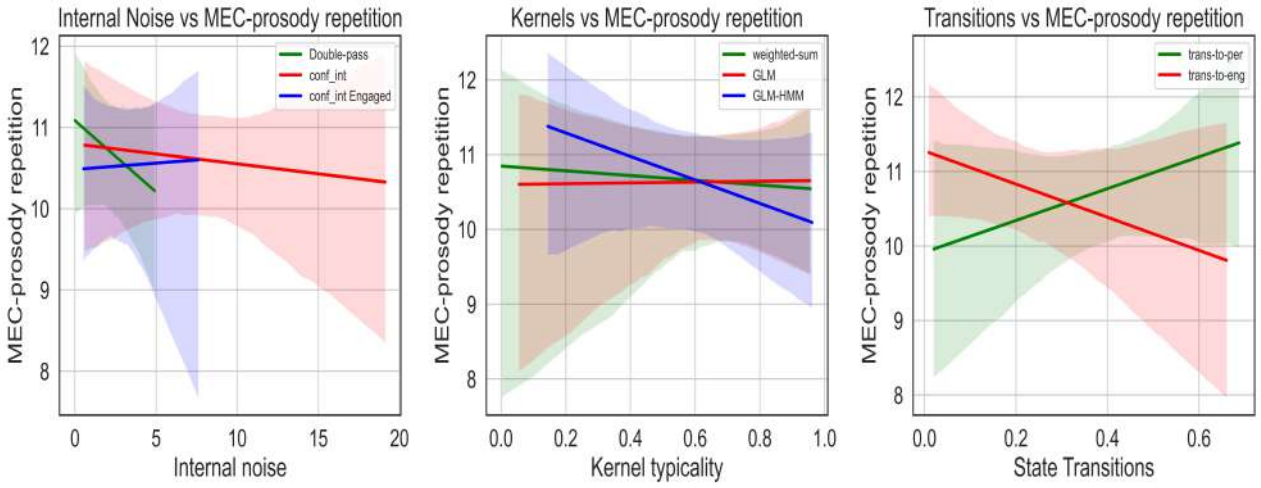


Fig. 8.17 **Relationship between behavioral metrics and prosody repetition task (MEC-repetition).** MEC-repetition (Y-axis, all panels) is scored out of 12. **(Left)** Internal noise: Double-pass (DP, old) and model-based (GLM, GLM-HMM) estimates all show non-significant relationships with MEC-repetition. **(Middle)** Kernel typicality (0–1): Neither weighted-sum (old) nor GLM-HMM (new, engaged trials only) kernel typicality measures are significantly associated with MEC-repetition. **(Right)** Transition probabilities (0–1, new): probability of transitioning to engaged or perseverative states shows no significant relationship with repetition performance.

GLM ($\beta = -0.20$, $p = 0.36$) were not significantly related, suggesting that after removing perseverative trials, internal noise no longer predicts auditory processing abilities.

For transition probabilities, higher transition to the engaged state ($\beta = 5.17$, $p = 0.51$) was positively associated with AIRTAC, suggesting that patients who return to engagement more frequently tend to perform better in auditory processing, though not significantly. Conversely, transition to perseveration ($\beta = 0.31$, $p = 0.95$) was slightly positively associated with AIRTAC but not meaningfully.

The above analysis can be separated between the duration and intensity discrimination subtasks of AIRTAC. For AIRTAC duration discrimination (Figure 8.19), kernel typicality across all three methods was significantly correlated with better discrimination ($p < 0.05$). Additionally, the old internal noise measure was negatively correlated ($\beta = -0.82$, $p = 0.007$), but this effect disappears in HMM ($\beta = -0.27$, $p = 0.45$), suggesting that the previous effect may have been driven by perseveration.

For AIRTAC intensity discrimination (Figure 8.20), only kernel typicality showed sig-

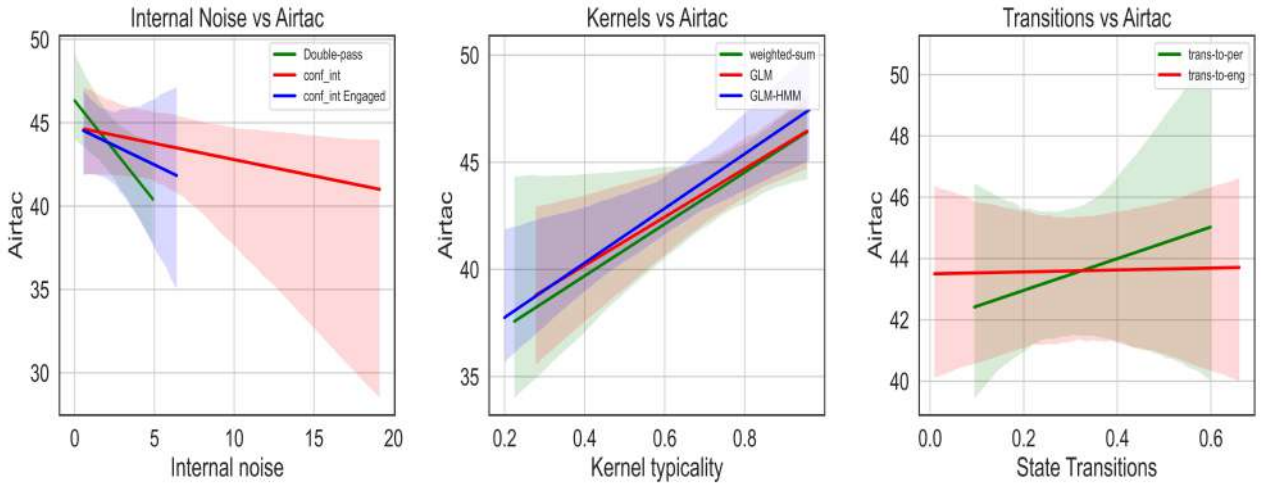


Fig. 8.18 **Relationship between behavioral metrics and central auditory processing (AIRTAC).** AIRTAC (Y-axis, all panels) measures auditory discrimination ability. **(Left)** Internal noise: The Double-pass (DP, old) estimate shows a significant negative association with AIRTAC, while model-based (GLM, GLM-HMM) internal noise estimates show no significant relationship. **(Middle)** Kernel typicality (0–1): Both weighted-sum (old) and GLM-HMM (new, engaged trials only) kernel typicality measures are strongly, positively associated with AIRTAC scores. **(Right)** Transition probabilities (0–1, new): Probability of transitioning to engaged or perseverative states shows no significant association with auditory processing abilities.

nificant correlations ($p < 0.05$), while internal noise estimates were not predictive. Transition probabilities were not significantly related, but transition to the engaged state ($\beta = 3.89$, $p = 0.39$) had a weak positive association, suggesting that patients who return to engagement more often may have slightly better intensity discrimination.

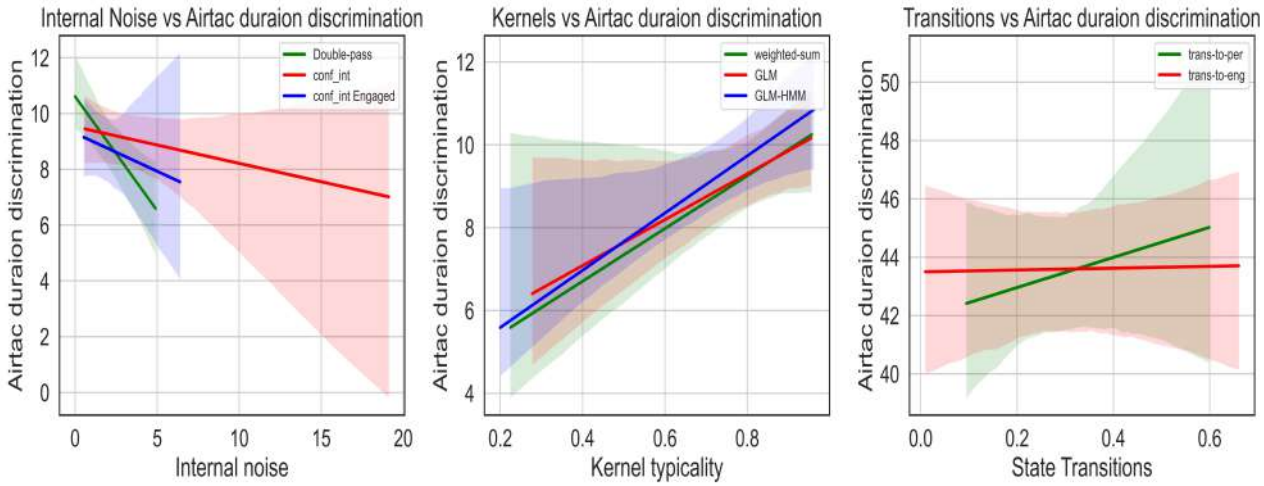


Fig. 8.19 **Relationship between behavioral metrics and AIRTAC duration discrimination.** The Y-axis in all panels represents AIRTAC duration discrimination performance. **(Left)** Internal noise: The Double-pass (DP, old) measure is significantly negatively associated with discrimination scores, while GLM and GLM-HMM (new, engaged trials) estimates show no significant association. **(Middle)** Kernel typicality (0–1): All methods (weighted-sum, GLM, GLM-HMM) show a significant positive relationship, indicating that higher kernel typicality is associated with better discrimination. **(Right)** Transition probabilities (0–1, new): No significant association is observed between state transition probabilities and AIRTAC duration discrimination.

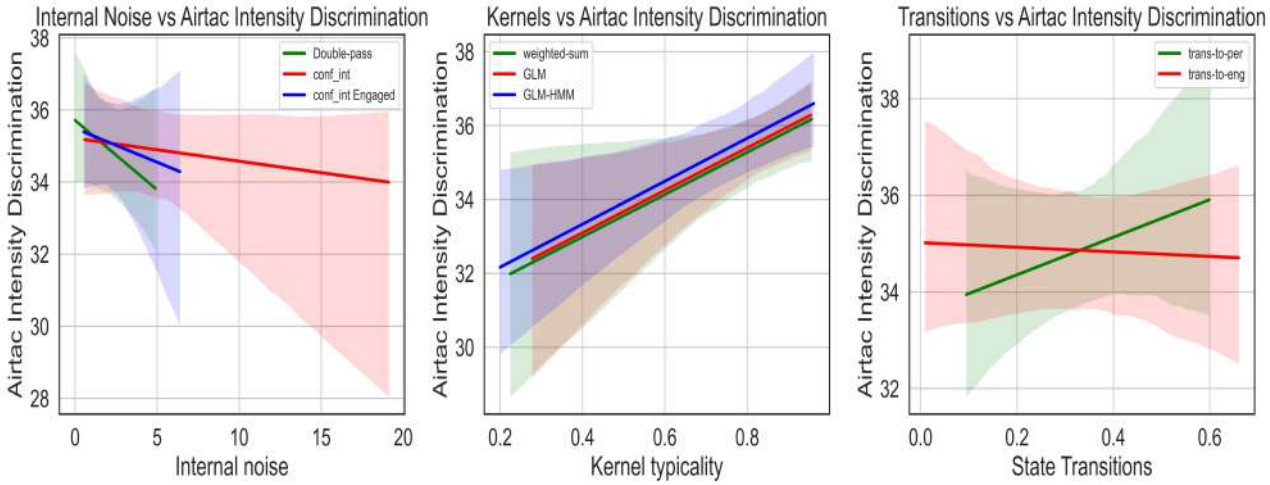


Fig. 8.20 **Relationship between behavioral metrics and AIRTAC intensity discrimination.** The Y-axis represents AIRTAC intensity discrimination performance. **(Left)** Internal noise (double-pass, GLM, and GLM-HMM) shows no significant association with discrimination scores. **(Middle)** Kernel typicality (weighted-sum, GLM, GLM-HMM) shows a significant positive correlation, indicating that greater kernel typicality is linked to better intensity discrimination. **(Right)** Transition probabilities (to engaged/perseverative states, new): neither measure is significantly related to AIRTAC intensity discrimination, though transitions to the engaged state show a weak positive trend.

8.4.3 Correlation with LAMA

LAMA is an index of auditory attentional control, reflecting a participant's ability to stay engaged in a stimulus-driven task. The regression results (Figure 8.21) show that internal noise from HMM was significantly negatively correlated with LAMA precision ($\beta = -0.21$, $p = 0.0004$). This suggests a link between attentional control and internal noise in engaged trials. Interestingly, the old internal noise measure was also negatively correlated but not significant ($\beta = -0.10$, $p = 0.24$), suggesting that the previously observed effect was hindered by perseveration.

Transition probabilities were not significantly related, but transition to the engaged state ($\beta = 0.34$, $p = 0.72$) showed a weak positive association, suggesting that attentional control may associate with the capacity to re-engage to the task.

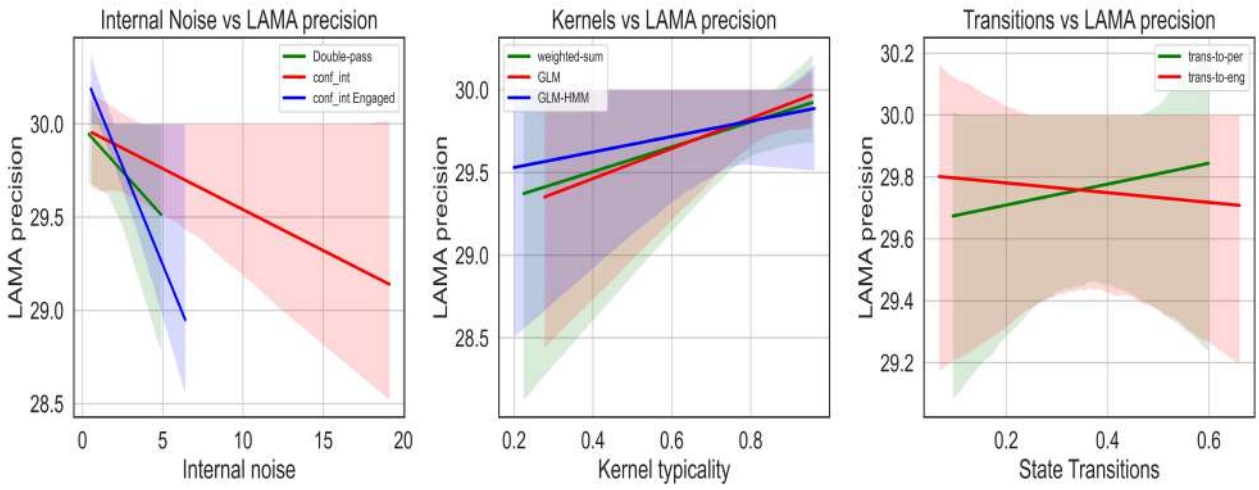


Fig. 8.21 **Relationship between behavioral metrics and LAMA precision (auditory attentional control).** Y-axis: LAMA precision score. **(Left)** Internal noise: GLM-HMM (new, engaged trials) shows a significant negative association with LAMA precision, indicating that lower internal noise is linked to better attentional control. Double-pass (old) and GLM (all trials) show weaker, non-significant associations. **(Middle)** Kernel typicality: positive, non-significant associations with LAMA precision for all kernel methods. **(Right)** Transition probabilities (model-derived): not significantly associated with LAMA precision, though transition to the engaged state shows a weak positive trend.

8.4.4 Correlations with MBEA

For MBEA (Montreal Battery of Evaluation of Amusia) total scores to estimate the impairment in music perception (Figure 8.22), no significant relationships were found with any predictors ($p > 0.05$).

The internal noise measured by double-pass had a weak positive relationship ($\beta = 0.05$, $p = 0.97$), which shifted to a negative trend in HMM ($\beta = -1.51$, $p = 0.45$), suggesting that perseveration may have contributed to false positives in prior estimates.

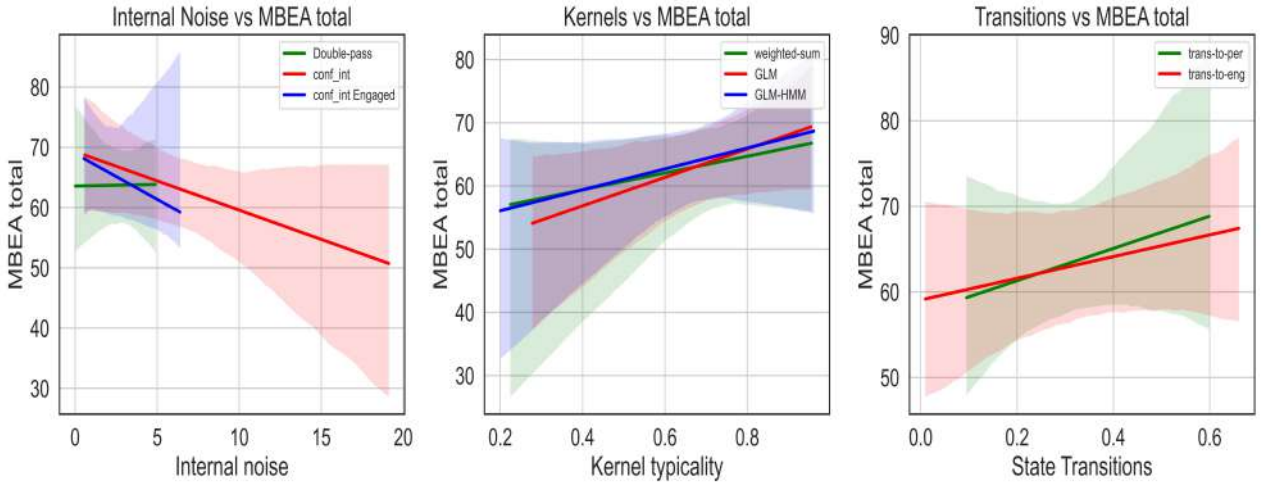


Fig. 8.22 **Relationship between behavioral metrics and MBEA total score (musical perception).** Y-axis: MBEA total (Montreal Battery of Evaluation of Amusia). **(Left)** Internal noise: Double-pass (old method) shows a weak positive (non-significant) association with MBEA, shifting to a non-significant negative trend for GLM-HMM (engaged trials only), suggesting prior effects may have been influenced by perseveration. **(Middle)** Kernel typicality: all methods show positive, non-significant relationships with MBEA total. **(Right)** Transition probabilities: probability of switching to engaged or perseverative states (GLM-HMM) shows weak, non-significant positive trends with MBEA total.

8.4.5 Correlations with HADS

For HADS (anxiety and depression scale, Figure 8.23), significant correlations were found with internal noise from HMM ($\beta = 2.28$, $p = 0.003$), GLM ($\beta = 0.95$, $p = 0.016$), and the double-pass method ($\beta = 2.30$, $p = 0.018$), as well as with kernel typicality estimated using HMM ($\beta = -16.31$, $p = 0.029$).

This suggests that higher internal noise is associated with increased psychological distress, indicating that patients with more perceptual instability may experience greater emotional difficulties, or conversely, that anxiety may cause perceptual inconsistencies.

Additionally, the significant correlation with kernel typicality from HMM suggests that after removing perseverative trials, patients with higher anxiety or depression symptoms may still exhibit altered stimulus-driven processing.

For transition probabilities, neither transition to perseveration ($\beta = -0.62$, $p = 0.95$) nor transition to the engaged state ($\beta = -12.80$, $p = 0.17$) were significantly related.

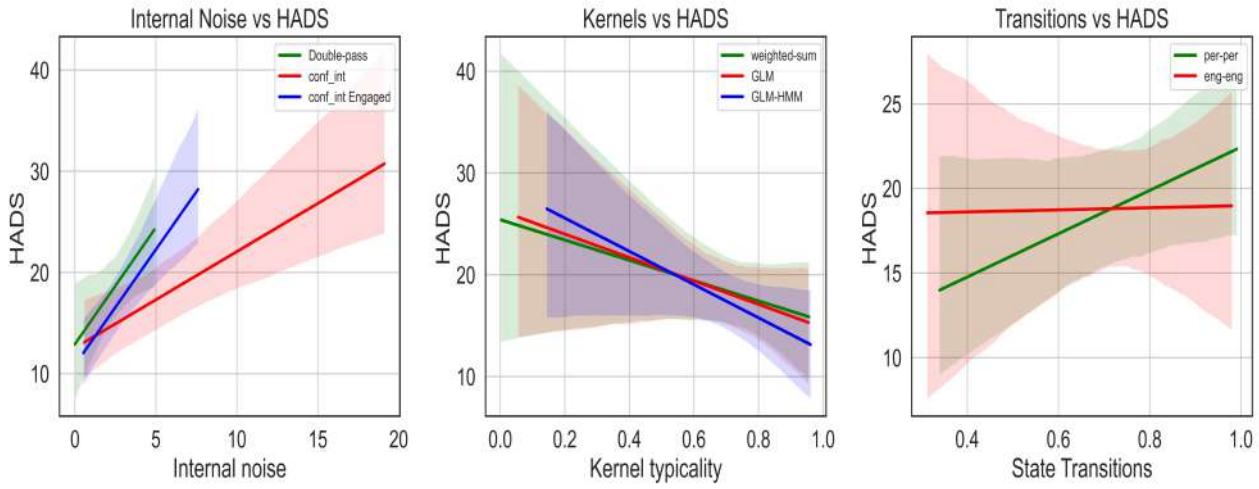


Fig. 8.23 **Relationship between behavioral metrics and HADS (Hospital Anxiety and Depression Scale).** Y-axis: HADS total score. **(Left)** Internal noise: All estimation methods (double-pass, GLM, GLM-HMM) show significant positive associations with HADS, indicating that greater perceptual instability is linked to increased psychological distress. **(Middle)** Kernel typicality: Kernel typicality estimated by GLM-HMM is significantly negatively correlated with HADS, suggesting that less typical (less control-like) kernels are associated with higher anxiety and depression symptoms; other methods show non-significant trends. **(Right)** Transition probabilities: Probability of switching to engaged or perseverative states (GLM-HMM) are not significantly related to HADS scores.

8.5 Conclusion

In this chapter, we re-analysed the patient data presented in Chapter 4, comparing reverse correlation estimates obtained with previous methods (kernel: weighted-sum/CI and GLM, internal noise: double-pass) and the new GLM-HMM method, which combines the identification of engaged trials, the estimation of kernel with the GLM method in the engaged state, the estimation of internal noise with the GLM confidence-interval method in the engaged state, and an additional estimate of transition probabilities between states.

Figure 8.24 summarizes the main differences found between methods.

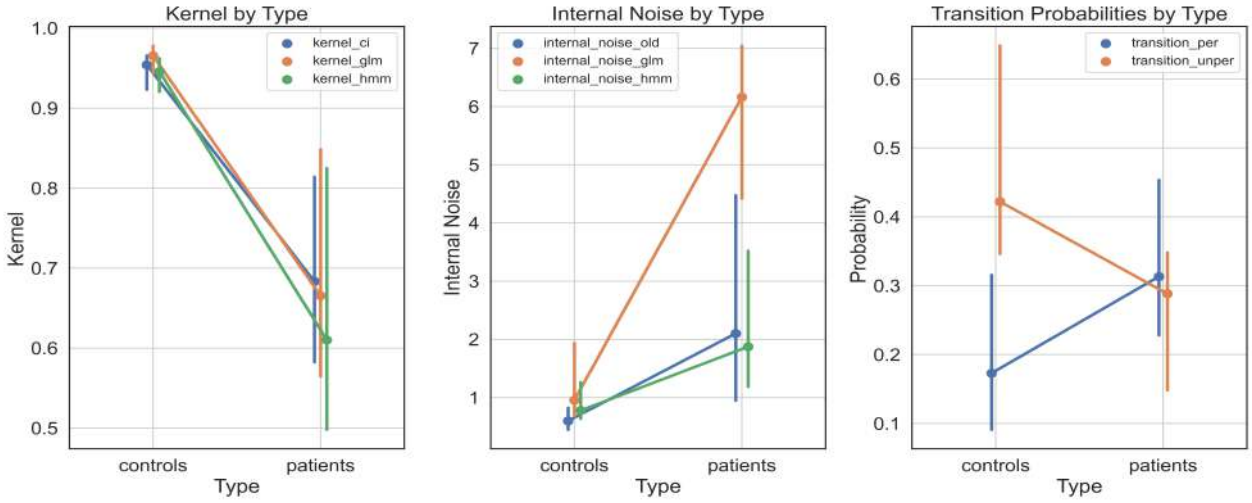


Fig. 8.24 **A comparison of reverse correlation parameter estimates across methods, for controls and patients. Left:** kernel typicality, as estimated with CI, GLM and GLM-HMM. **Middle:** Internal noise, as estimated with double-pass, GLM confidence intervals and GLM-HMM confidence intervals. **Right:** Transition probabilities estimated with the GLM-HMM method.

First, the comparison of different methods for measuring kernel (CI, GLM, and GLM-HMM) revealed no significant differences across methods (Figure 8.24-left), and how these estimates associate with clinical measures (Table 8.1). Methodologically, this suggests that the impact of perseveration on kernel estimation was limited and did not impact the clinical interpretation of kernel typicality. Clinically, while HMM-HLM kernel typicality allowed differentiating between controls and patients, including controls and asymptomatic (MEC>9) patients, it did not correlate significantly with MEC within the control group. As in Chapter 4, GLM-HMM kernel typicality was positively associated with AIR-

TAC and not with LAMA and MBEA. It was also found negatively associated with HADS, while that association was not significant with the old estimates in Chapter 4 (Table 8.1).

Tab. 8.1 Regression results between biomarkers (1): Kernel estimates

Outcome	Kernel CI	Kernel GLM	Kernel HMM
MEC	2.9308 (0.3999)	3.3431 (0.3271)	-1.7669 (0.6138)
Prosodie C	3.2471 (0.1792)	3.2887 (0.1657)	-0.1950 (0.9372)
Prosodie R	-0.3163 (0.8326)	0.0545 (0.9705)	-1.5720 (0.2866)
Airtac	12.0803 (0.0095) *	11.2837 (0.0075) *	12.7224 (0.0031) *
Airtac Dur Discr	6.3606 (0.0170) *	5.5400 (0.0249) *	6.9076 (0.0048) *
Airtac Int Discr	5.7197 (0.0498) *	5.7437 (0.0274) *	5.8148 (0.0362) *
Lama Prec	0.7515 (0.2393)	0.9095 (0.1137)	0.4689 (0.4596)
MBEA Total	13.2686 (0.4384)	22.5102 (0.1349)	16.5472 (0.3111)
HADS	-9.9711 (0.2012)	-11.5011 (0.1305)	-16.3155 (0.0297) *
Transition to PER	-0.1189 (0.3749)	-0.0974 (0.4525)	-0.2030 (0.1055)
Transition to ENG	0.4184 (0.0083) *	0.4221 (0.0057) *	0.4213 (0.0047) *

Second, the methods used to estimate internal noise showed clear differences (Figure 8.24-middle). After removing perseverative trials in GLM-HMM, internal noise decreased significantly in both (MEC>9 and MEC$\leq$ 9) patient groups, showing that a large portion of the high noise estimates made with the double-pass method was likely inflated by perseverative responses rather than true perceptual variability. Clinically, tighter GLM-HMM noise estimates still allowed differentiating MEC$\leq$ 9 patients and controls, as well as MEC$\leq$ 9 and MEC >9 patients. However, it did not separate MEC > 9 and controls contrary to old estimates in Chapter 4, suggesting that this difference was driven by estimation errors. Unsurprisingly, because of these large differences, internal noise correlations with clinical measures were quite different from the original results of Chapter 4. While double-pass noise estimates were associated with MEC-comprehension and AIRTAC, new GLM-HMM estimates aren't. Conversely, while double-pass noise estimates were not associated with LAMA, new estimated are (Table 8.2). Taken together, this suggests that GLM-HMM internal noise estimates are in fact quite validly distinct

from auditory processing measures (MEC and AIRTAC), and associated with attentional control (LAMA), but that the correct identification of these trends was hindered by estimation errors in our previous results. Additionally, the positive relation between internal noise and HADS was preserved across methods.

Tab. 8.2 Regression results between biomarkers (2): internal noise estimates

Outcome	Internal Noise DP	Internal Noise GLM	Internal Noise HMM
MEC	-0.7989 (0.0709)	-0.2573 (0.1598)	-0.2284 (0.6035)
Prosodie C	-0.6229 (0.0433) *	-0.2327 (0.0660)	-0.2440 (0.4328)
Prosodie R	-0.1761 (0.3650)	-0.0245 (0.7587)	0.0155 (0.9330)
Airtac	-1.2038 (0.0374) *	-0.1952 (0.3582)	-0.4606 (0.4987)
Airtac Dur Discr	-0.8179 (0.0071) *	-0.1316 (0.2639)	-0.2728 (0.4484)
Airtac Int Discr	-0.3859 (0.2916)	-0.0637 (0.6152)	-0.1878 (0.6297)
Lama Prec	-0.0957 (0.2379)	-0.0439 (0.0962)	-0.2105 (0.0004) *
MBEA Total	0.0549 (0.9786)	-0.9706 (0.1411)	-1.5146 (0.4484)
HADS	2.3045 (0.0180) *	0.9513 (0.0162) *	2.2825 (0.0033) *
Transition to PER	-0.0015 (0.9345)	0.0068 (0.3490)	0.0036 (0.8235)
Transition to ENG	-0.0405 (0.0611)	-0.0293 (0.0004) *	-0.0253 (0.2267)

Finally, regarding state transitions (Figure 8.24-right), we did not observe a significant difference between controls and patients in their probability of transitioning into the perseverative state. This indicates that perseveration is not exclusive to patients; controls can also exhibit repetitive, non-stimulus-driven responses. However, the critical distinction was in the ability to exit the perseverative state and return to the engaged state. Patients showed a greater tendency to remain stuck in perseveration, struggling to return to an engaged, stimulus-driven behavior, whereas controls were able to transition back more easily, maintaining a flexible and adaptive response pattern. These findings suggest that the primary deficit in patients is not an increased tendency to enter perseveration but rather an impaired ability to recover from it. This inability to return to an engaged state may be a key factor in the observed difficulties in stimulus-driven perception and auditory attention, reinforcing the importance of distinguishing engaged trials

from perseverative ones when assessing perceptual and cognitive flexibility. Clinically, transition probabilities did not associate statistically with any of the reported clinical measures (Table 8.3), which we can either take to suggest that transition probabilities are an idiosyncratic feature of the task which doesn't have clinical relevance or (more optimistically) that perseveration dynamics are, in fact a blind spot of traditional clinical measures and an important, understudied symptom of post-stroke rehabilitation. The next and final chapter of this thesis will expand on this perspective by analysing some of the properties of patient perseveration revealed by the GLM-HMM analysis.

Tab. 8.3 Regression results between biomarkers (3): transition probabilities

Outcome	Transition to PER	Transition to ENG
MEC	5.9660 (0.1809)	1.0443 (0.8053)
Prosodie C	3.8327 (0.2240)	3.2653 (0.2665)
Prosodie R	2.1333 (0.2649)	-2.2210 (0.2099)
Airtac	5.1727 (0.5068)	0.3143 (0.9516)
Airtac Dur Discr	1.2847 (0.7702)	0.7895 (0.7848)
Airtac Int Discr	3.8880 (0.3926)	-0.4752 (0.8758)
Lama Prec	0.3372 (0.7276)	-0.1572 (0.8271)
MBEA Total	18.8525 (0.4484)	12.6752 (0.4379)
HADS	-0.6187 (0.9523)	-12.7998 (0.1728)
Transition to PER	-	0.0036 (0.8236)
Transition to ENG	-0.0253 (0.2267)	-

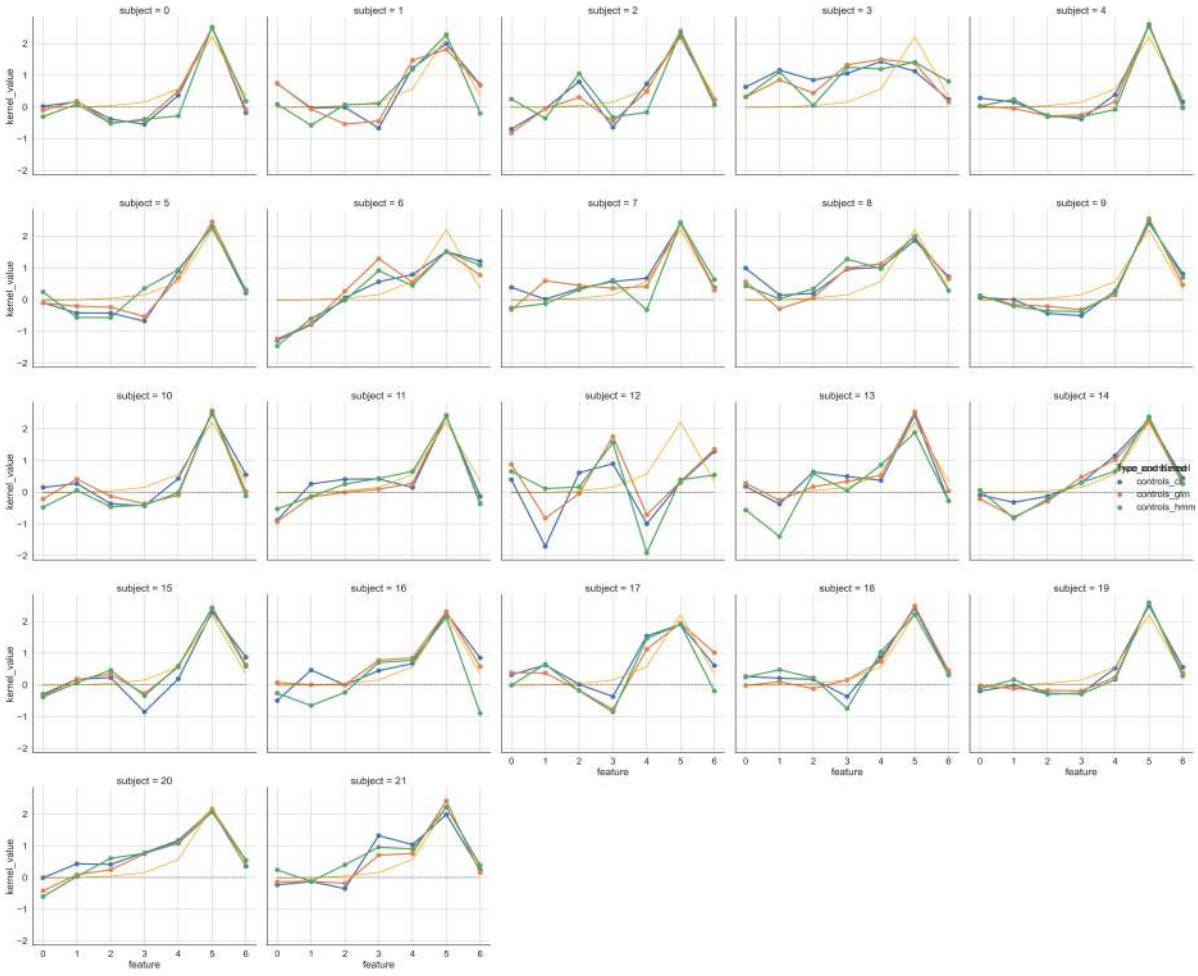


Fig. 8.25 **Kernel estimation using 3 methods for each healthy participant** Normalized kernel weights for pitch discrimination features across 22 control participants, extracted using three different modeling approaches: Classification images(CI), standard GLM and GLM-HMM

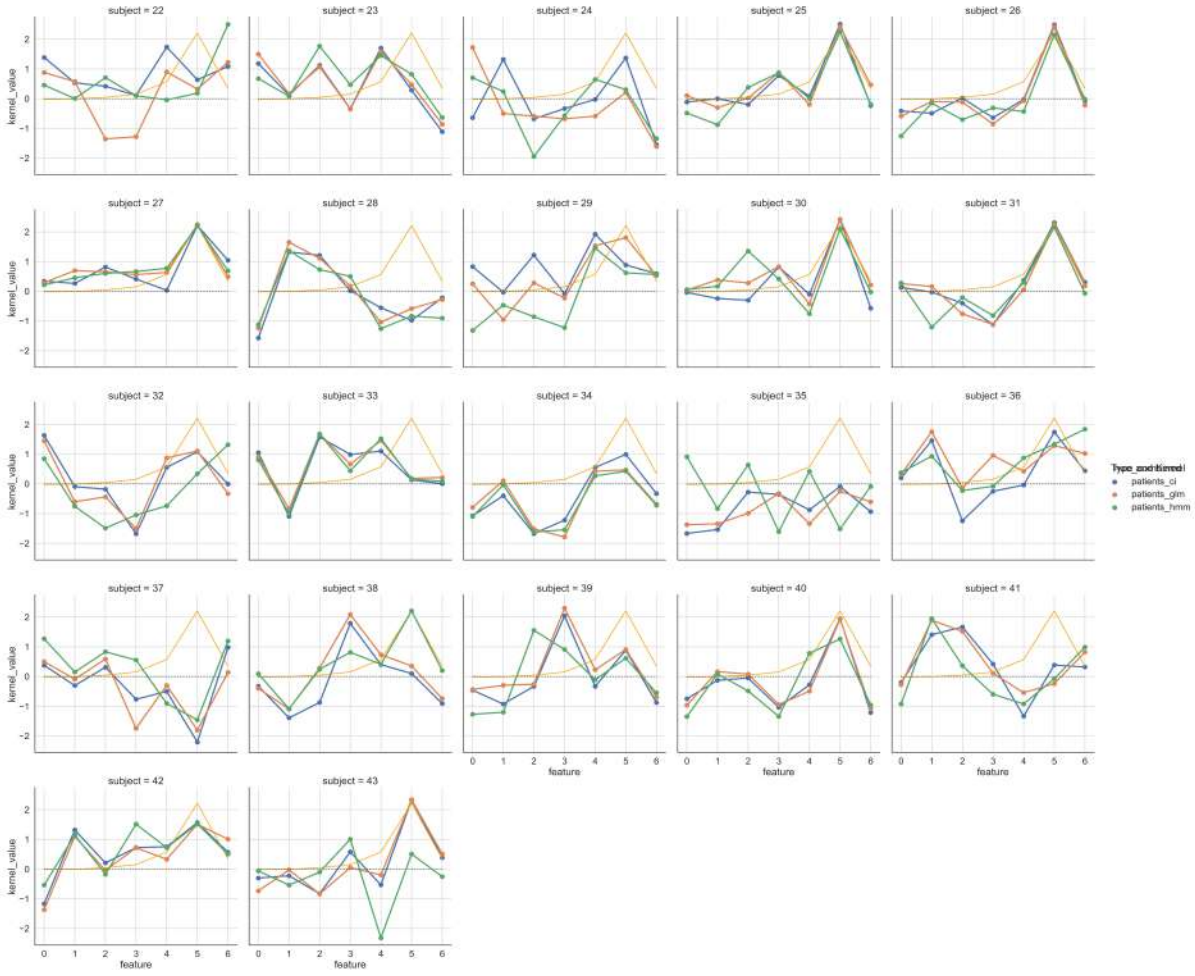


Fig. 8.26 **Kernel estimation using 3 methods for each RH stroke patient** Normalized kernel weights for pitch discrimination features across 22 RH stroke patients, extracted using three different modeling approaches: Classification images(CI), standard GLM and GLM-HMM

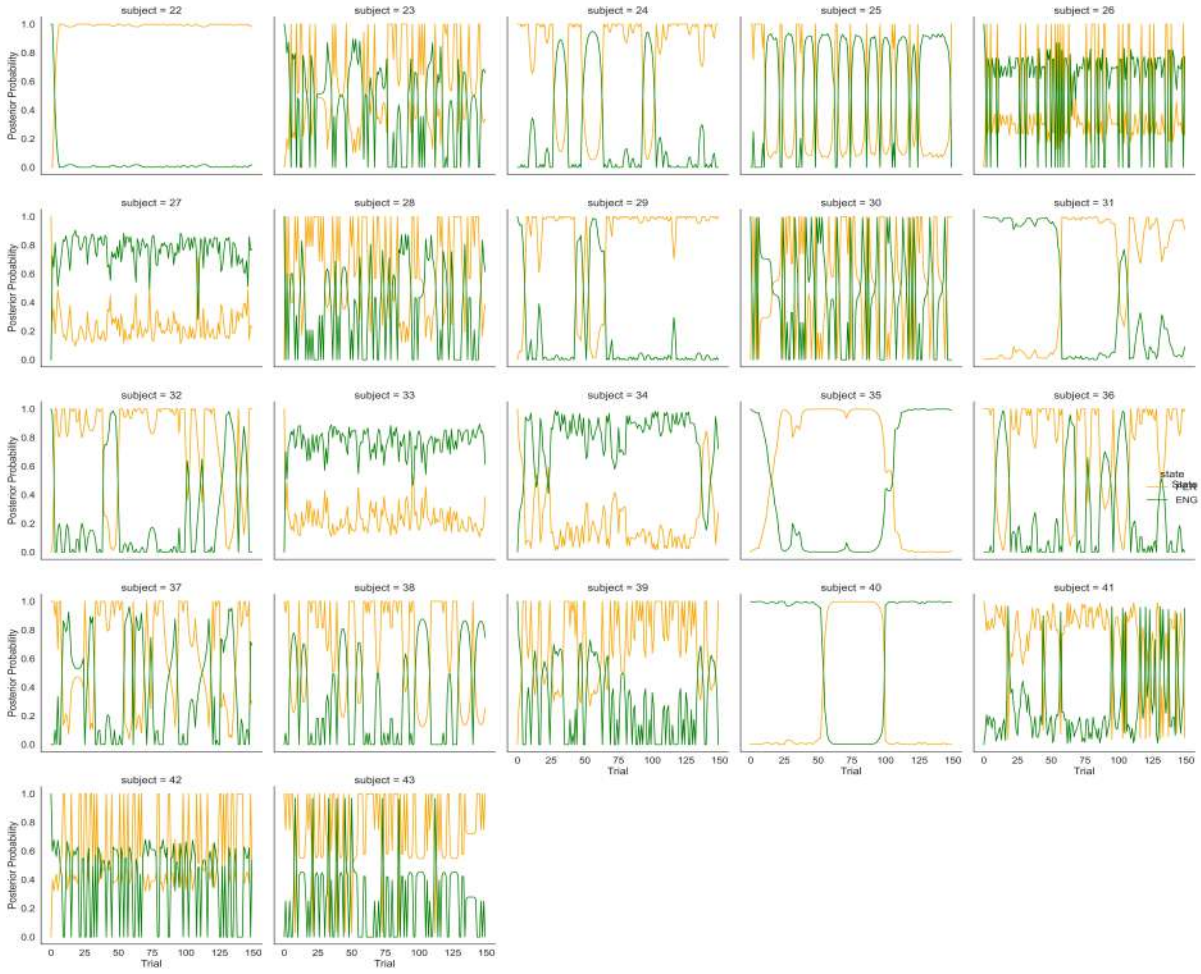


Fig. 8.27 **Posterior state probabilities 2-state GLM-HMM across trials for RH stroke patients** Posterior probabilities of the perseverative (PER, orange) and engaged (ENG, green) states over trials for each patient extracted using the GLM-HMM model.

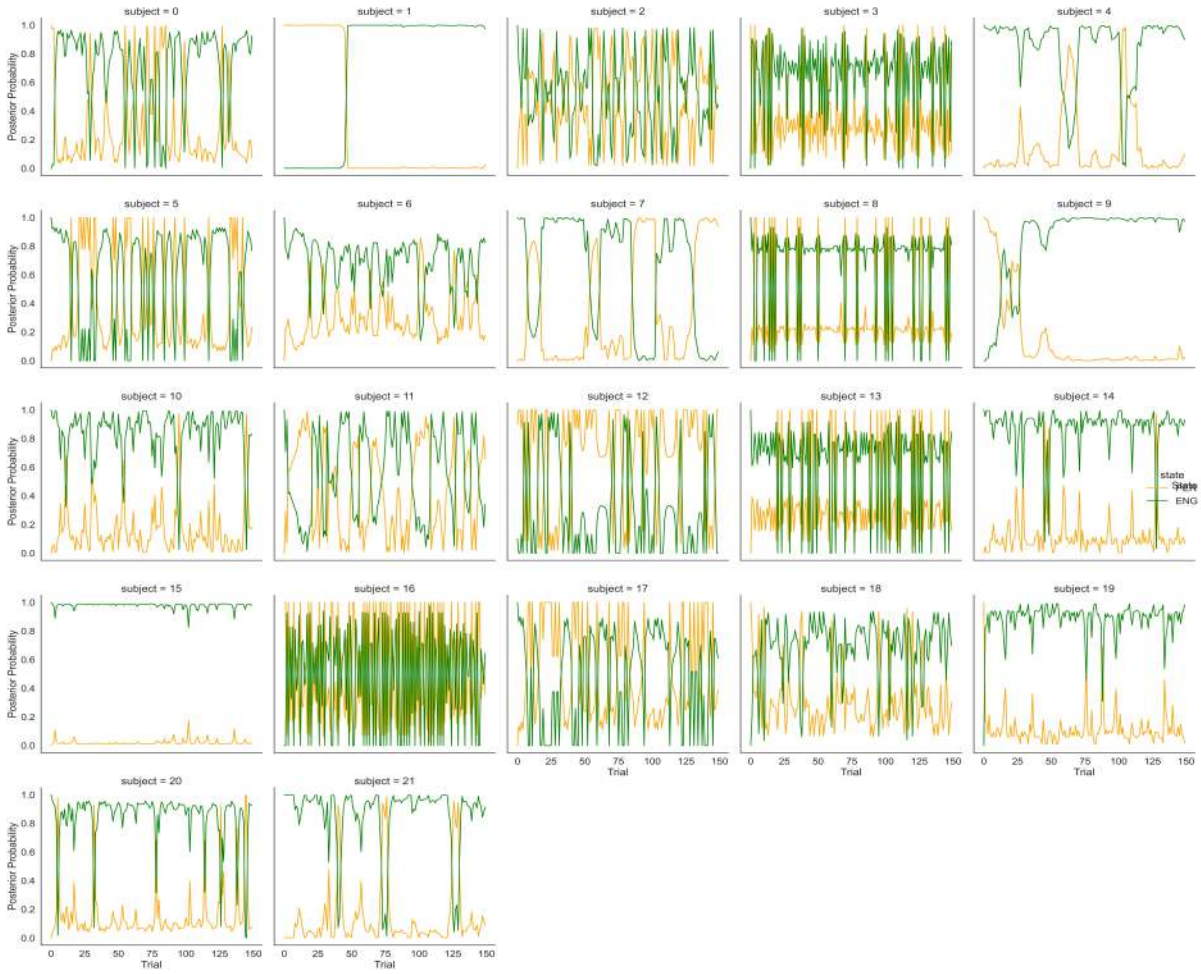


Fig. 8.28 **Posterior state probabilities of 2-state GLM-HMM across trials for healthy participants**

CHAPTER 9

Perseveration analysis

The reanalysis of patient data with the new methods developed in this thesis (GLM-HMM) has confirmed that patients and controls differ in the tendency to perseverate, and provided new ways to quantify this tendency using transition probabilities. It was shown, in particular, that patients do not so much differ from controls in the probability to perseverate, but rather in their ability to exit the perseverative state and return to the engaged state. Clinically, transition probabilities did not associate statistically with any of the reported clinical measures, which may suggest that perseveration is an under-studied and under-measured aspect of post-stroke symptoms.

In this chapter, we provide an exploratory analysis of the characteristics of perseveration in our sample, and in particular whether we can link state switching to certain characteristics of the stimuli or responses in the vicinity of the switch. To do do, we exploit the fact that fitting the GLM-HMM model provides labels (i.e., posterior state probabilities) for the most probable state at each trial, and that the characteristics of these trials (stimuli, responses, response times) are random but known.

Our goal is to explore whether specific parameters influence the likelihood of patients entering the perseverative state and, conversely, how their behavior differs when they

do not perseverate. Additionally, we investigate whether there are notable differences between patients and controls in this posterior analysis, providing further insights into the mechanisms underlying perseveration and engagement in stimulus-driven tasks.

9.1 Response times in the vicinity of state switching

The data collected as part of Chapter 4 includes a measure of response time (RT), i.e., the time for each trial between the end of the (automated) play of the second sound stimuli and when the response button is pressed.

We examine here how RT differs when participants remain in the same state versus when they transition between states, potentially reflecting shifts in processing dynamics or attentional fluctuations. While RT is often interpreted as an indirect marker of cognitive processing demands or decision uncertainty, it can be influenced by a range of factors beyond cognitive effort alone (Wylie, Yao, Sandry, & DeLuca, 2021).

We first analyse RTs by state. Controls exhibit minimal differences in reaction time between the perseverative and engaged states (ENG=1.43s, PER=1.47s), suggesting a relatively stable processing mechanism. Patients not only show longer RTs overall but also slightly slower responses in the perseverative state (PER=2.18s) than the engaged state (ENG=2.07s). This pattern of results suggest that the GLM-HMM PER state may not be subtended by identical cognitive states in controls and patients. In particular, the slow down of response time in patient perseveration suggests that it is not an effortless, disengaged process.

To further document such differences, we then compared RTs between engaged trials that either remained engaged (i.e., all trials in engaged sequences except trials that immediately preceded a switch to perseveration) and engaged trials that immediately preceded a switch to perseveration. When transitioning from engagement to perseveration, controls demonstrate a slight decrease of RT decrease (1.45 to 1.25). Conversely, patients exhibit a significant increase of RT at the time of the switch (1.95 to 2.65), which is consistent with overall longer PER trials seen above. Again, this suggests that switching to PER is not the same process for controls and patients. While the former may simply passively

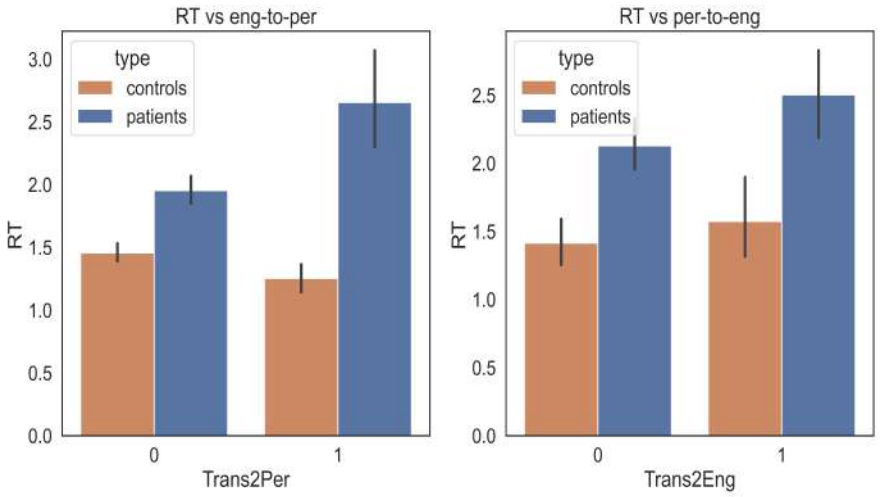


Fig. 9.1 **Reaction times (RT) associated with state transitions in controls and patients. Left:** RTs for engaged trials which either remain engaged (trans2per = 0) or immediately precede a switch to perseveration ($Trans2Per = 1$). **Right:** RTs for perseverated trials which either remain perseverating ($Trans2Eng = 0$), or immediately preceded a switch back to the engaged state (Trans2Eng=1). Controls show minimal RT changes across state transitions, whereas patients exhibit notably higher RTs overall and particularly during transitions, especially when re-engaging with the stimulus. Error bars indicate a 95% confidence interval.

and effortlessly disengage from the task, the latter appear to engage in more effortful processing despite the repetitive response.

Examining transitions from perseveration back to engagement (trans2eng), we observe that RTs increase compared to non-transitioning PER trials for both groups. PER trials at the end of a PER sequence show a moderate slowdown in controls (1.41 to 1.57) and a more pronounced slowdown in patients (2.12 to 2.5). This could suggest that re-engaging with the stimulus after a period of perseveration requires additional cognitive processing, particularly for patients who struggle more to shift back to a stimulus-driven mode.

Taken together, this analysis of RTs highlights potential differences in cognitive flexibility between controls and patients. While controls can fluidly transition between states, with minimal changes of RTs, patients show increased reaction times when shifting in either direction, reflecting both difficulty in maintaining engagement and challenges in exiting the perseverative state. More generally, PER trials in patients are associated with slower RTs, which suggest that patient perseveration is not an effortless process of

disengagement from the task but rather involves effortful processing.

9.2 Are perseverations driven by sequences of identical responses ?

While perseveration by definition involve sequences of repeated identical responses, we first ask whether there is any evidence in the data that repeated responses may also *precede* a switch to perseveration, which would be consistent with the view that perseverating participants have difficulties inhibiting a response once it is established.

To do so, we extracted all trials in engaged state that immediately preceded a switch to perseveration (i.e., the last trial in every sequence of successive engaged trials), and computed the empirical distribution of the number of consecutive identical responses that was ongoing at these trials, before the switch. Because all trials are random, the mode of that distribution is likely equal to 1, but it may be larger either because of a un/lucky sequence of stimuli or response tendencies. We then compared this empirical distribution with the expected theoretical distribution, which is the cumulated density of a binomial distribution with probability $p = 0.5$. We compared both distributions with the Kolmogorov–Smirnov test, a nonparametric test of the equality of one-dimensional probability distributions.

In Figure 9.2, we compare the empirical distribution of identically repeated responses before a perseveration switch for control participants and the theoretical (expected) distribution. There was no statistical difference between the two distributions (K-S test,$p = 0.88$), suggesting that switch to perseveration-like states in controls do not statistically coincide with preceding sequences of repeated responses.

We conducted the same analysis for patients (Figure 9.3. The empirical distribution had a larger tail, with some switches occurring after 10 or more successive identical responses. That pattern over the group of patients differed statistically from the expected distribution (K-S, $p = 0.0036$). This suggests that, before a switch to perseveration, patients exhibit more repetitive responses than what would be expected with random

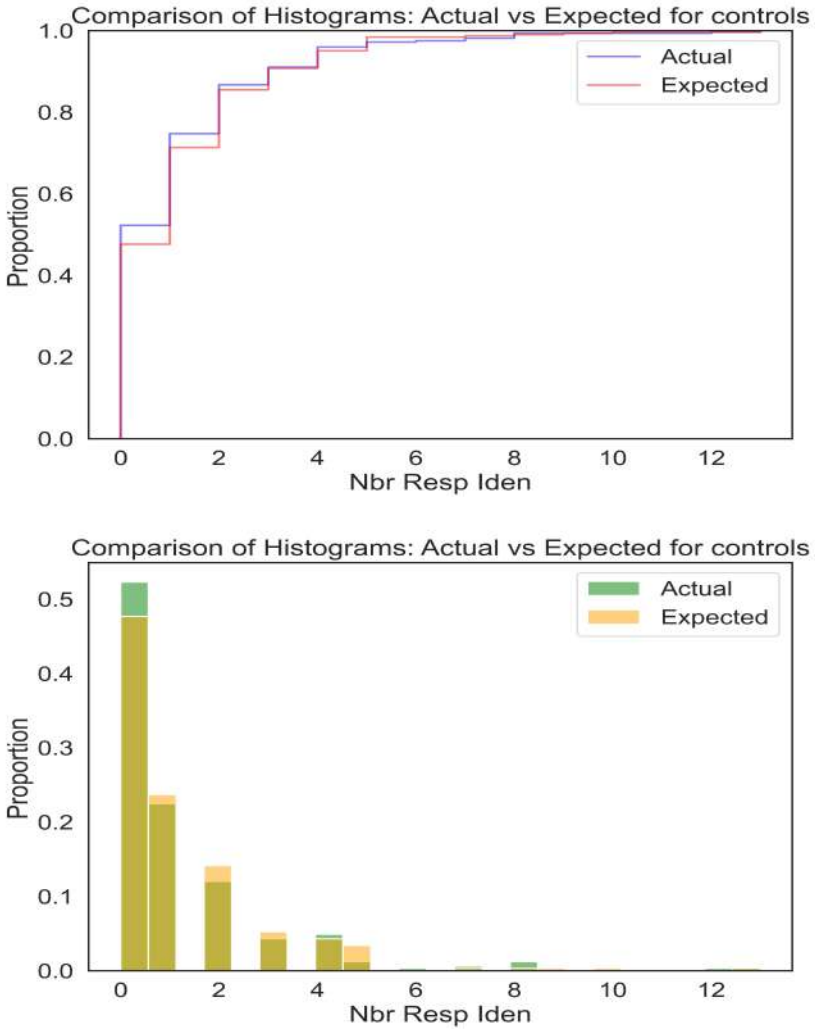


Fig. 9.2 **Comparison of the expected and observed distribution of consecutive identical responses before a perseveration switch for controls. Top** shows the cumulative distributions. **Bottom** shows the corresponding histogram proportions. In engaged trials that immediately precede a perseveration switch, controls do not exhibit excessive repetition beyond what is predicted by stimulus-driven behavior (K-S test, $p = 0.88$).

stimuli.

In conclusion, we find statistical evidence that, in patients, switches to perseveration tend to be preceded with longer sequences of identical responses than what would be expected with a binomial response strategy. These responses may be legit, i.e., due to a un/lucky streak of trials that warrant identical responses, or indicate that perseverating behaviour has already started before the labeled switch by the GLM-HMM.

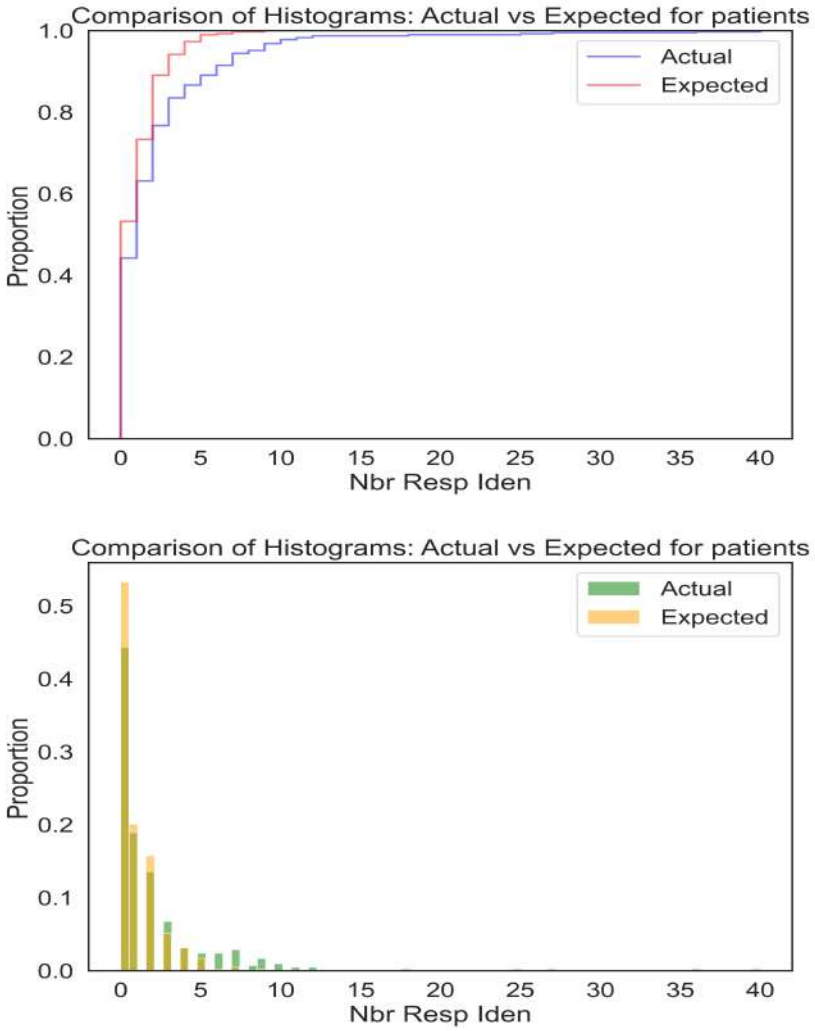


Fig. 9.3 **Comparison of the expected and observed distribution of consecutive identical responses before a perseveration switch for patients. Top:** shows the cumulative distributions. **Bottom:** shows the corresponding histogram proportions. In contrast to controls, patients exhibit a significant excess of repeated responses before a switch, beyond what is expected (K-S test, $p = 0.0036$).

9.3 Are perseverations driven by trial difficulty ?

Another potential driver for state switching is trial difficulty, measured as the alignment between the stimulus difference and the participant's engaged kernel. Because transition to perseveration may indicate disengagement from the task, it is possible that such switch coincide with trials for which responses are more difficult, for instance, trials in which the two stimuli are relatively similar or differ along directions that are not relevant to the

participant's kernel.

To operationalize this, trial clarity was quantified as the absolute value of the dot product between the stimulus difference on each trial and the engaged-state kernel weights (similar to trial weighting in Chapter 6). Lower values correspond to more ambiguous and challenging trials, while higher values indicate that the stimulus was well aligned with the participant's internal template, making the trial easier to respond to.

We then compared the distributions of trial difficulty between engaged trials that remained engaged (i.e., all trials in engaged sequences except trials that immediately preceded a switch to perseveration) and engaged trials that immediately preceded a switch to perseveration, using a Kolmogorov–Smirnov test. Figure 9.4-right, compares the distribution of trial difficulty in both types of engaged trials, for patients. The results show no significant difference in trial difficulty between trials where participants transitioned to perseveration and those where they remained engaged (K-S, $p = 0.67$). This suggests that perseveration is not directly driven by stimulus difficulty. If trial difficulty were a key factor, we would expect participants to transition into perseveration more frequently in harder trials. However, since no such effect is observed, it implies that internal cognitive processes, response biases, or fatigue may play a greater role in state transitions than trial difficulty alone.

Figure 9.4-left shows the distribution of trial difficulty in patient trials starting in the perseverative state, comparing those that transitioned to engaged with those that did not. Here, a significant difference in trial difficulty is observed (K-S, $p = 0.005$), with the PER $\longrightarrow$ ENG trial displaying more stimulus-kernel alignment (i.e., lower difficulty). To confirm this effect, a logistic regression was fitted to model the probability of transitioning from the perseverative to the engaged state as a function of trial difficulty for patients. The results revealed a small but statistically significant effect: easier trials were associated with a higher probability of transitioning to the engaged state (GLM, $p = 0.034$). This suggest that, while the transition to perseveration is not driven by difficult trials, stimulus clarity may help participants re-engage with the task following periods of perseveration.

We conducted the same analysis for controls (Figure 9.5). There was no significant difference in trial difficulty between trials where participants transitioned to perseveration

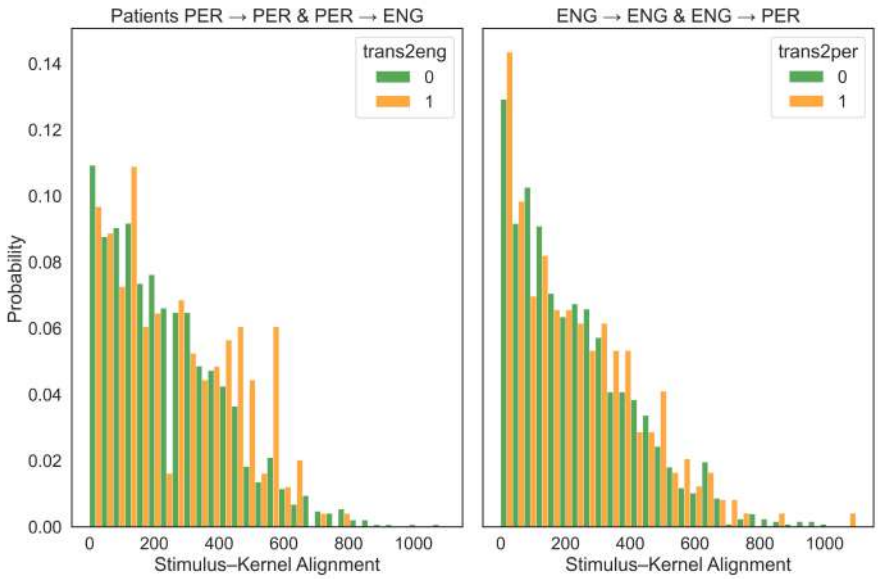


Fig. 9.4 **Distribution of stimulus alignment across transitions in patients. Left:** Perseverative trials that either transitioned to the engaged state (orange) or remained perseverative (green). A significant difference was found here, suggesting that in lower trial difficulty participants can re-engage to the task. **Right:** Engaged trials that either transitioned to the perseverative state (orange) or remained engaged (green). No significant difference was observed between the two distributions.

and those where they remained engaged (K–S, $p = 0.33$), and no significant difference in trial difficulty between perseverating trials where participants remained perseverating and those where they transitioned back to engagement (K–S, $p = 0.69$). Equivalently, stimulus alignment was not a significant predictor of state transitions in controls based on GLM analysis. This suggests that, contrary to patients, trial difficulty does not influence state transitions in healthy participants.

9.4 Conclusion

In this relatively short final chapter, we reported on an exploratory analysis aiming to clarify the stimulus and response characteristics in the vicinity of state change and, in particular whether the switch to PER is partly stimulus-driven.

Our results suggest that the PER state does not necessarily involve the same cognitive processes in controls and patients. For the former, transition to PER is associated with

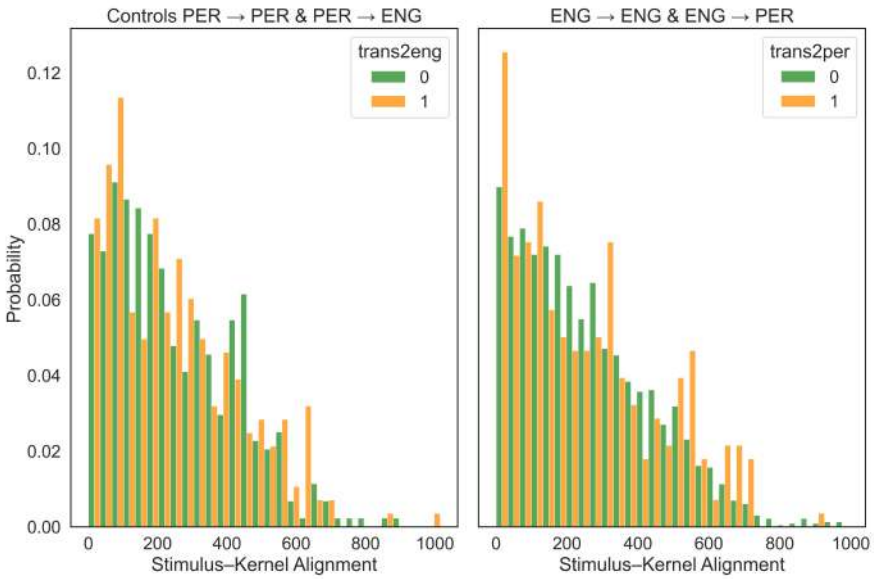


Fig. 9.5 **Distribution of stimulus alignment across transitions in controls. Left:** Perseverative trials that either transitioned to the engaged state (orange) or remained perseverative (green). No significant difference was observed between the two distributions **Right:** Engaged trials that either transitioned to the perseverative state (orange) or remained engaged (green). No significant difference was observed between the two distributions.

faster RTs, suggesting that it corresponds to effortless disengagement from the task. For patients however, transitions to PER is associated with a marked slowdown of responses, which suggests that they engage in more effortful processing despite the repetitive response.

We then examined whether changes of states are associated with particular characteristics of the stimuli and responses that precede the change. In patients, switches to perseveration tend to be preceded with longer sequences of identical responses than what would be expected with a binomial response strategy, and switches back to engagement coincide with smaller trial difficulty. No such effects were found in controls.

While they are only correlational, such effects are interesting for several reasons. First, to test for a possible causal effect, one could imagine artificially introducing sequences of trials that warrant identical responses and seeing if these make patients more likely to enter perseveration; conversely, when patients are found to be perseverating, one could imagine a closed-loop procedure where they are presented easier trials to test if this drive them back to task engagement. Second, these effects may suggest a possible strategy to

avoid perseverating behaviour in patients, by avoiding to prevent long sequences of trials that warrant identical responses. Further experimental work will be needed to test the causality of this possible relation.

Finally, the fact that controls do not display such associations confirms the results of response time and suggest that PER states labeled by the GLM-HMM in control data do not correspond to the same underlying cognitive process than patients.

CHAPTER 10

Conclusion

10.1 Summary of contributions

The goal of this thesis was to improve the diagnosis and comprehension of deficits of prosody perception after a brain stroke by capitalizing on a recently developed psychophysical technique, reverse correlation.

As discussed in **Chapters 2** and **3**, survivors of right-hemisphere stroke may experience deficits in prosody perception that persist from the acute to chronic phase. These impairments can often remain subtle and undetected during speech therapy sessions when assessed using standard batteries. Reverse correlation appears to be a promising method to uncover the psychophysical parameters of such deficits when present.

In **Chapter 4**, we presented an initial analysis of reverse correlation data, collected prior to the thesis (the work of speech therapy students Mélissa Jeulin, Pauline Bardet, Pauline Commère, supervised by Marie Villain, Emmanuel Ponsot and JJ Aucouturier), using the reverse correlation paradigm of Ponsot, Burred, et al. (2018), and the two classical methods of classification-image kernel estimation and double-pass internal noise estimation. Our results showed that analyzing the patients' kernels and internal noise

provide a way to identify distinct pathological profiles and uncover the diverse sensory and cognitive mechanisms that underlie prosody processing impairments after stroke. These results were published in Scientific Reports (2024).

Based on this data, in **Chapter 5**, we identified two key challenges that affect the reliability of classical methods for estimating internal noise and mental representations: the limited number of trials and perseveration. The limited trial count imposes constraints on the precision of estimations, particularly that of internal noise. Perseveration disrupts both the stimulus-response relation and response variability, leading to large estimation errors in both kernel and noise. Taken together, the results of Chapter 4 and 5 have led to the definition of a problem statement for the remainder of the thesis: in order to apply reverse correlation to patient populations and to fatiguable/perseverating stroke patients in particular, we needed to develop kernel and internal noise estimation methods that are both robust to a low number of trials and to local disruptions of decision strategy such as perseveration.

The next part of this thesis has introduced several methodological contributions that address this problem. In **Chapter 6**, we introduced and evaluated three new methods to estimate internal noise in the absence of double-pass trials. These methods address both problems above: first, because they do not rely on double-pass data, they allow using the complete set of trials in an experiment to infer internal noise, which is likely an advantage with small-trial setups such as here. Second, because they do not restrict internal noise measurements to specific blocks, they are also expected to be less sensitive to local perturbations such as perseverations. Of these three methods, two were developed by collaborators (Ladislas Nalborczyk, JJ Aucouturier and Marie Villain) and one (GLM confidence intervals) by myself; my contribution is also to compare them against one another and against the double-pass method. Perhaps most importantly for our context here, at low/very-low number of trials (ex. $n = 100$ single-pass, or $n = 50$ repeated trials), we found that the double-pass method achieves an unimpressive 48% relative error, while we evaluated our best methods at an error of 30%, and even 20% when increasing the number of trials to $n = 600$. This chapter was presented in the format of a preprint, co-written with L.N, JJ.A. and M.V, and which is intended for submission at a

methodological journal such as *Behaviour Research Methods* or *Quantitative Methods for Psychology*.

In **Chapter 7**, we then introduced a new method to conjointly estimate both linear-observer parameters and perseverating episodes, using a joint model with two latent states (input-output hidden Markov model, or GLM-HMM). We showed that this model is able to recover perseverating episodes by taking into account not only repeated responses but also stimuli-response relations and therefore, to improve the accuracy of kernel and noise estimates across non-perseverated episodes.

In part IV, we then applied this new and improved set of methods to reanalyse the experimental data of Chapter 4 and confirm or infirm the clinical interpretability of reverse correlation estimates. In **Chapter 8**, we found that the comparison of different methods for measuring kernel (CI, GLM, and GLM-HMM) revealed no significant differences across methods and how these estimates associate with clinical measures. However, the methods used to estimate internal noise showed clear differences. After removing perseverative trials in GLM-HMM, internal noise decreased significantly in both (MEC>9 and MEC<9) patient groups, showing that a large portion of the high noise estimates made with the double-pass method was likely inflated by perseverative responses. Clinically, contrary to double-pass estimates, GLM-HMM noise estimates were no longer correlated with MEC-comprehension and AIRTAC but were correlated to LAMA. Finally, GLM-HMM estimates of transition probabilities between states provided another set of potential biomarkers and showed that the primary deficit in patients is not an increased tendency to enter perseveration but rather an impaired ability to recover from it.

In a short final chapter (**Chapter 9**), we used the GLM-HMM's ability to estimate perseverating episodes to provide a supplementary analysis of what factors may influence perseveration in patients. Our results suggest, first, that the perseverating state does not necessarily involve the same cognitive processes in controls and patients. For the former, transition to PER is associated with faster RTs, suggesting that it corresponds to effortless disengagement from the task. For patients, however transitions to PER is associated with a marked slowdown of responses, which suggests that they engage in more effortful processing despite the repetitive response. Second, we examined whether changes

of states are associated with particular characteristics of the stimuli and responses that precede the change. We found that, in patients, switches to perseveration tend to be preceded with longer sequences of identical responses than what would be expected with a binomial response strategy, and switches back to engagement coincide with smaller trial difficulty or stimulus-kernel alignment.

10.2 Perspectives

This thesis offers initial insights into how reverse-correlation parameters, when correctly estimated, can be used to inform the clinical evaluation of brain stroke patients suffering perceptual deficits. Naturally, many questions remain open for future research.

From a methodological perspective, one of the most compelling directions involves developing joint neural embedding and behavioral models, also investigating physiological (EEG, eye-tracking, etc.) markers for attention detection (Cai, Su, Xie, & Li, 2021), or consciousness (described in Shadlen and Kiani (2011) as a "decision to engage"). Another particularly valuable approach would be to design a closed-loop stimulus presentation system. Such a system would adapt in real time to transitions in mental state and task difficulty while also allowing for significant reductions in experimental time. This would be especially beneficial for participants prone to fatigue and would facilitate the rapid and precise extraction of mental representations.

From a cognitive perspective, one of the most interesting perspective of this work is that reverse-correlation, and more generally perceptual decision-making models generally fail to account for the possibility that individuals may transition between different mental states multiple times within a single session. Although the study includes a limited sample of 22 right-hemisphere post-stroke patients and 22 healthy participants, it effectively captures the interplay between internal experiences and external stimuli. The findings suggest that individuals process external events differently depending on their internal state. Such states may be a various natures. Here, perseveration and engagement may be described as "attentional". It would also be interesting to explore other forms of internal states, for instance affective states related to task performance, such as discouragement

or pride. Another open question is the neural bases of such state changes. Recent studies about mind wandering (Xie & Xu, 2024; Zhang & Kool, 2024), have emphasized the importance of investigating the interconnectivity of brain networks, cerebral activity patterns across regions, and the cognitive functions associated with shifting mental states.

Finally, from a clinical perspective, while this thesis has focused on data recorded on a sample of RH stroke patients, the same paradigm could easily be used with other populations – perhaps most obviously LH stroke patients. Comparing RH and LH stroke patients on their reverse correlation parameters would inform on the specificity of kernel vs internal noise alterations and speak to the debate of the lateralization of prosodic processing. More generally, given a larger sample of patients, it would be interesting to perform symptom-lesion mapping to investigate whether kernel, noise and perseveration parameters are associated with lesions in different sub-networks. In addition, these psychophysical markers could be studied longitudinally, e.g., in patients undergoing brain tumor resection or during post-stroke rehabilitation, to assess how perceptual and cognitive parameters evolve over time. Finally, because the reverse correlation task can be adapted to a variety of auditory/linguistic judgements (e.g., emotional prosody), the models developed in this thesis hold promise not only for stroke patients but also for individuals with ASD, schizophrenia, and other brain injuries associated with perceptual or emotional atypicalities.

Bibliography

Abbey, C. K., Eckstein, M. P., & Bochud, F. O. (1999). Estimation of human-observer templates in two-alternative forced-choice experiments. In *Medical imaging 1999: Image perception and performance* (Vol. 3663, pp. 284–295).

Abbruzzese, M., Ferri, S., & Scarone, S. (1996). Performance on the wisconsin card sorting test in schizophrenia: Perseveration in clinical subtypes. *Psychiatry Research*, *64*, 27-33.

Adl Zarrabi, A., Jeulin, M., Bardet, P., Commère, P., Naccache, L., Aucouturier, J.-J., ... Villain, M. (2024). A simple psychophysical procedure separates representational and noise components in impairments of speech prosody perception after right-hemisphere stroke. *Scientific Reports*, *14*(1), 15194.

Ahumada, A., & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751–1756.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of vision*, *2 1*, 121-31.

Angeleri, R., Bosco, F. M., Zettin, M., Sacco, K., Colle, L., & Bara, B. G. (2008). Communicative impairment in traumatic brain injury: A complete pragmatic assessment. *Brain and Language*, *107*, 229-245.

Anikin, A. (2020). The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics*, *29*(2), 226–247.

Ashwood, Z. C. (2022). *Probabilistic models for characterizing animal learning and decision-making* (Unpublished doctoral dissertation). Princeton, NJ : Princeton University.

Ashwood, Z. C., Roy, N. A., Stone, I. R., Urai, A. E., Churchland, A. K., Pouget, A., & Pillow, J. W. (2020). Mice alternate between discrete strategies during perceptual decision-making. *bioRxiv*.

Banuazizi, A., & Creswell, C. (1999). Is that a real question? final rises, final falls, and discourse function in yes-no question intonation. *CLS*, *35*, 1–14.

Barker-Collo, S., Feigin, V., Lawes, C., Senior, H., & Parag, V. (2010). Natural history of attention deficits and their influence on functional recovery from acute stages to 6 months after stroke. *Neuroepidemiology*, *35*(4), 255–262.

Baum, S. R., & Dwivedi, V. (2003). Sensitivity to prosodic structure in left- and right-hemisphere-damaged individuals. *Brain and Language*, *87*, 278-289.

Baum, S. R., & Pell, M. D. (1999). The neural bases of prosody: Insights from lesion studies and neuroimaging. *Aphasiology*, *13*(8), 581–608.

Behrens, S. J. (1989). Characterizing sentence intonation in a right hemisphere-damaged population. *Brain and Language*, *37*, 181-200.

Belyk, M., & Brown, S. (2014a). Perception of affective and linguistic prosody: an ale meta-analysis of neuroimaging studies. *Social cognitive and affective neuroscience*, *9 9*, 1395-403.

Belyk, M., & Brown, S. (2014b). Perception of affective and linguistic prosody: an ale meta-analysis of neuroimaging studies. *Social cognitive and affective neuroscience*, *9*(9), 1395–1403.

Benedetti, V., Weill-Chounlamountry, A., Pradat-Diehl, P., & Villain, M. (2022). Assessment tools and rehabilitation treatments for aprosodia following acquired brain injury: A scoping review. *International Journal of Language & Communication Disorders*, *57*(3), 474–496.

Bengio, Y., & Frasconi, P. (1996). Input-output hmms for sequence processing. *IEEE Transactions on Neural Networks*, *7*(5), 1231–1249.

Benton, E. C., & Bryan, K. (1996). Right cerebral hemisphere damage: incidence of language problems. *International Journal of Rehabilitation Research*, *19*, 47–54.

Blake, M. L., Duffy, J. R., Myers, P. S., & Tompkins, C. A. (2002). Prevalence and patterns of right hemisphere cognitive/communicative deficits: Retrospective data

from an inpatient rehabilitation unit. *Aphasiology*, *16*(4-6), 537–547.

Blumstein, S. E., & Cooper, W. E. (1974). Hemispheric processing of intonation contours. *Cortex; a journal devoted to the study of the nervous system and behavior*, *10 2*, 146-58.

Bolinger, D. L. M. (1958). A theory of pitch accent in english. *WORD*, *14*, 109-149.

Borod, J., Welkowitz, J., & Obler, L. (1992). The new york emotion battery. *Unpublished materials, Mount Sinai Medical Center, Department of Neurology, New York*.

Bowers, D., Blonder, L., & Heilman, K. (1998). *Florida affect battery*. Center for Neuropsychological Studies, Department of Neurology Florida, USA.

Brady, M. C., Godwin, J., Enderby, P., Kelly, H., & Campbell, P. (2016). Speech and language therapy for aphasia after stroke: An updated systematic review and meta-analyses. *Stroke*, *47*(10), e236–e237.

Brosnan, M. B., Dockree, P. M., Harty, S., Pearce, D., Levenstein, J. M., Gilbert, C., ... Demeyere, N. (2020). Lost in time: Temporal monitoring elicits clinical decrements in sustained attention post-stroke. *Journal of the International Neuropsychological Society*, *28*, 249 - 257.

Bunker, L. D., & Hillis, A. E. (2022). Vascular syndromes: Revisiting classification of poststroke aphasia. *Handbook of clinical neurology*, *185*, 37–55.

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. iv. observer inconsistency. *Journal of the Optical Society of America A*, *5*(4), 617–627.

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, *214*(4516), 93–94.

Burred, J. J., Ponsot, E., Goupil, L., Liuni, M., & Aucouturier, J. (2018). Cleese: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PLoS ONE*, *14*.

Cai, S., Su, E., Xie, L., & Li, H. (2021). Eeg-based auditory attention detection via frequency and channel neural attention. *IEEE Transactions on Human-Machine Systems*, *52*(2), 256–266.

Cancelliere, A. E. B., & Kertesz, A. (1990). Lesion localization in acquired deficits of emotional expression and comprehension. *Brain and Cognition*, *13*, 133-147.

Caulfield, M. D., Chen, P., Barry, M. M., & Barrett, A. (2017). Which perseverative behaviors are symptoms of spatial neglect? *Brain and cognition*, *113*, 93–101.

Commandeur, J. J., & Koopman, S. J. (2007). *An introduction to state space time series analysis*. Oxford university press.

Côté, H., Payer, M., Giroux, F., & Joanette, Y. (2007). Towards a description of clinical communication impairment profiles following right-hemisphere damage. *Aphasiology*, *21*(6-8), 739–749.

Cramer, S. C., Richards, L. G., Bernhardt, J., & Duncan, P. (2023). Cognitive deficits after stroke. *Stroke*, *54*(1), 5–9.

Crane, T., & Patterson, S. (2012). *History of the mind-body problem*. Routledge.

Cumming, T. B., Brodtmann, A., Darby, D., & Bernhardt, J. (2012). Cutting a long story short: reaction times in acute stroke are associated with longer term cognitive outcomes. *Journal of the neurological sciences*, *322*(1-2), 102–106.

Cumming, T. B., Packer, M., Kramer, S. F., & English, C. (2016). The prevalence of fatigue after stroke: a systematic review and meta-analysis. *International Journal of stroke*, *11*(9), 968–977.

Cummins, D. (2000). A memorable patient: Fascinating rhythm. *BMJ*, *320*, 575-575.

Davis, G. A., O'Neil-Pirozzi, T. M., & Coon, M. (1997). Referential cohesion and logical coherence of narration after right hemisphere stroke. *Brain and language*, *56*(2), 183–210.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

dong Wang, X., Wang, M., & Chen, L. (2013). Hemispheric lateralization for early auditory processing of lexical tones: Dependence on pitch level and pitch contour. *Neuropsychologia*, *51*, 2238-2244.

Eggermont, J. J. (1993). Wiener and volterra analyses applied to the auditory system. *Hearing Research*, *66*, 177-201.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*, 86 - 88.

Eling, P. (2011). Lichtheim's golden shot. *Cortex*, *47*(4), 501–508.

Ernst, E. (1990). A review of stroke rehabilitation and physiotherapy. *Stroke*, *21*(7), 1081–1085.

Etchepare, A., & Prouteau, A. (2018). Toward a two-dimensional model of social cognition in clinical neuropsychology: a systematic review of factor structure studies. *Journal of the International Neuropsychological Society*, *24*(4), 391–404.

Fahrmeir, L., Tutz, G., Fahrmeir, L., & Tutz, G. (2001). State space and hidden markov models. *Multivariate Statistical Modelling Based on Generalized Linear Models*, 331–383.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, *9*(4), 292–303.

Falmagne, J.-C. (1985). Elements of psychophysical theory..

Fechner, G. T. (1948). Elements of psychophysics, 1860.

Feigin, V. L., Barker-Collo, S., Parag, V., Senior, H., Lawes, C., Ratnasabapathy, Y., ... Group, A. S. (2010). Auckland stroke outcomes study: Part 1: Gender, stroke types, ethnicity, and functional outcomes 5 years poststroke. *Neurology*, *75*(18), 1597–1607.

Flowers, H. L., Skoretz, S. A., Silver, F. L., Rochon, E., Fang, J., Flamand-Roze, C., & Martino, R. (2016). Poststroke aphasia frequency, recovery, and outcomes: a systematic review and meta-analysis. *Archives of physical medicine and rehabilitation*, *97*(12), 2188–2201.

Frazier, P. I. (2018). A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Freedman, M., Black, S. E., Ebert, P., & Binns, M. A. (1998). Orbitofrontal function, object alternation and perseveration. *Cerebral cortex*, *8 1*, 18-27.

Friedman, H. S., Prince, L. M., Riggio, R. E., Dimatteo, M. R., Casella, D., Devlin, P., ... Friedman (1980). Understanding and assessing nonverbal expressiveness: The affective communication test. *Journal of Personality and Social Psychology*, *39*, 333-351.

Friedman, S., Samuelian, J.-C., Lancrenon, S., Even, C., & Chiarelli, P. (2001). Three-

dimensional structure of the hospital anxiety and depression scale in a large french primary care population suffering from major depression. *Psychiatry research*, *104*(3), 247–257.

Gabet, A., Béjot, Y., Touzé, E., Woimant, F., Suissa, L., Grave, C., ... Olié, V. (2024). Epidemiology of stroke in france. *Archives of Cardiovascular Diseases*, *117*(12), 682–692.

Gandola, M., Toraldo, A., Invernizzi, P., Corrado, L., Sberna, M., Santilli, I., ... Paulesu, E. (2013). How many forms of perseveration? evidence from cancellation tasks in right hemisphere patients. *Neuropsychologia*, *51*(14), 2960–2975.

Gandour, J., Tong, Y., Wong, D., Talavage, T. M., Dzemidzic, M., Xu, Y., ... Lowe, M. J. (2004). Hemispheric roles in the perception of speech prosody. *NeuroImage*, *23*, 344-357.

Gillespie, D. C., Bowen, A., & Foster, J. K. (2006). Memory impairment following right hemisphere stroke: a comparative meta-analytic and narrative review. *The Clinical Neuropsychologist*, *20*(1), 59–75.

Goh, H.-T., & Stewart, J. C. (2019). Poststroke fatigue is related to motor and cognitive performance: A secondary analysis. *Journal of Neurologic Physical Therapy*, *Publish Ahead of Print*.

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021a). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, *12*(1), 861.

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021b). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, *12*(1), 861.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics..

Hamilton, R. H., Chrysikou, E. G., & Coslett, B. (2011). Mechanisms of aphasia recovery after stroke and the role of noninvasive brain stimulation. *Brain and language*, *118*(1-2), 40–50.

Hanley, J. A., et al. (1989). Receiver operating characteristic (roc) methodology: the state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307–335.

Hasan, B. A. S., Joosten, E., & Neri, P. (2012). Estimation of internal noise using double passes: Does it matter how the second pass is delivered? *Vision research*, *69*, 1–9.

Hausen, M., Torppa, R., Salmela, V. R., Vainio, M., & Särkämö, T. (2013). Music and speech prosody: a common rhythm. *Frontiers in Psychology*, *4*.

Heath, R. L., & Blonder, L. X. (2005). Spontaneous humor among right hemisphere stroke survivors. *Brain and Language*, *93*(3), 267–276.

Heilman, K. M., & Abell, T. V. D. (1980). Right hemisphere dominance for attention: the mechanism underlying hemispheric asymmetries of inattention (neglect). *Neurology*, *30*(3), 327–327.

Hellerman, J. (2003). Ann k. wennerstrom, the music of everyday speech: Prosody and discourse analysis. new york: Oxford university press, 2001. pp. xix, 317, pb. $24.95. *Language in Society*, *32*, 592 - 595.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67–99.

Howie, J. M. (1976). *Acoustical studies of mandarin vowels and tones* (Vol. 18). Cambridge University Press.

Hyndman, D., & Ashburn, A. (2003). People with stroke living in the community: Attention deficits, balance, adl ability and falls. *Disability and rehabilitation*, *25*(15), 817–822.

Hyndman, D., Pickering, R. M., & Ashburn, A. (2008). The influence of attention deficits on functional recovery post stroke during the first 12 months after discharge from hospital. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(6), 656–663.

Jin, Y.-P., Di Legge, S., Ostbye, T., Feightner, J. W., & Hachinski, V. (2006). The reciprocal risks of stroke and cognitive impairment in an elderly population. *Alzheimer's & dementia*, *2*(3), 171–178.

Joanette, Y., Ska, B., & Côté, H. (2004). Protocole montréal d'évaluation de la communication (protocole mec). *Isbergues, France: Ortho Édition*.

Johns, P. (2014). Clinical neuroscience: an illustrated colour text. *(No Title)*.

Johnson, K. O., Hsiao, S. S., & Yoshioka, T. (2002). Review: Neural coding and the basic law of psychophysics. *The Neuroscientist*, *8*, 111 - 121.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*, 455-492.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2019). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*, 237 - 265.

Kemmerer, D. (2022). *Cognitive neuroscience of language*. Routledge.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16), 10–10.

Kreiman, J., Gerratt, B. R., & Gabelman, B. (2002). Jitter, shimmer, and noise in pathological voice quality perception. *Journal of the Acoustical Society of America*, *112*(5), 2446.

Lancker, D. R. V. (1980). Cerebral lateralization of pitch cues in the linguistic signal. *Research on Language and Social Interaction*, *13*, 201-277.

Lancker, D. R. V., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: all errors are not created equal. *Journal of speech and hearing research*, *35 5*, 963-70.

Landrigan, J.-F., Zhang, F., & Mirman, D. (2021). A data-driven approach to post-stroke aphasia classification and lesion-based prediction. *Brain*, *144*(5), 1372–1383.

Langhorne, P., Bernhardt, J., & Kwakkel, G. (2011). Stroke rehabilitation. *The Lancet*, *377*(9778), 1693–1702.

Lecoffre, C., de Peretti, C., Gabet, A., Grimaud, O., Woimant, F., Giroud, M., ... Olié, V. (2017). L'accident vasculaire cérébral en france: patients hospitalisés pour avc en 2014 et évolutions 2008-2014. *Bulletin Epidémiologique Hebdomadaire-BEH*.

Lee, Y., & Schetzen, M. (1965). Measurement of the wiener kernels of a non-linear system by cross-correlation. *International Journal of Control*, *2*(3), 237–254.

Li, K., & Malhotra, P. A. (2015). Spatial neglect. *Practical neurology*, *15*(5), 333–339.

Licht, S. H. (1975). Stroke and its rehabilitation. *(No Title)*.

Lu, Z.-L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychological*

review, *115*(1), 44.

Maeshima, S., & Osawa, A. (2021). Memory impairment due to stroke.

Marmarelis, P. Z., & Marmarelis, V. Z. (2011). Analysis of physiological systems: The white-noise approach..

McCullagh, P. (2019). *Generalized linear models*. Routledge.

McLachlan, G. J., & Krishnan, T. (2008). *The em algorithm and extensions*. John Wiley & Sons.

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature neuroscience*, *11*(1), 103–107.

Mineault, P. J., Barthelmé, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of vision*, *9 10*, 17.1-24.

Minga, J., Sheppard, S. M., Johnson, M., Hewetson, R., Cornwell, P., & Blake, M. L. (2023). Apragmatism: The renewal of a label for communication disorders associated with right hemisphere brain damage. *International journal of language & communication disorders*, *58*(2), 651–666.

Müller, M. (2011). Generalized linear models. In *Handbook of computational statistics: Concepts and methods* (pp. 681–709). Springer.

Murray, R. F. (2011). Classification images: A review. *Journal of vision*, *11*(5), 2–2.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums. *Journal of vision*, *2 1*, 79-104.

Neri, P. (2010). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review*, *17*(6), 802–808.

Neri, P. (2020). Optimal templates for signal extraction by noisy ideal detectors and human observers. *Journal of Computational Neuroscience*, *49*, 1 - 20.

Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature*, *401*(6754), 695–698.

Nys, G., Zandvoort, M. v., Worp, H. v. d., Kappelle, L., & Haan, E. d. (2006). Neuropsychological and neuroanatomical correlates of perseverative responses in subacute stroke. *Brain*, *129*(8), 2148–2157.

Okazawa, G., Sha, L., Purcell, B. A., & Kiani, R. (2018). Psychophysical reverse correla-

tion reflects both sensory and decision-making processes. *Nature communications*, *9*(1), 3479.

Oliveira, C. R. d., Pagliarin, K. C., Calvette, L. d. F., Gindri, G., Argimon, I. I. d. L., & Fonseca, R. P. (2015). Depressive signs and cognitive performance in patients with a right hemisphere stroke. In *Codas* (Vol. 27, pp. 452–457).

Osses, A., Spinelli, E., Meunier, F., Gaudrain, E., & Varnet, L. (2023). Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation. *JASA express letters*, *3 9*.

O'Sullivan, M. J., Brownsett, S. L. E., & Copland, D. A. (2019). Language and language disorders: neuroscience to clinical practice. *Practical Neurology*, *19*, 380 - 388.

Paciaroni, M., & Acciarresi, M. (2019). Poststroke fatigue. *Stroke*.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O. T., & Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends in ecology & evolution*, *23 2*, 87-94.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. *Annals of the New York Academy of Sciences*, *999*.

Pessiglione, M., Blain, B., Wiehler, A., & Naik, S. (2025). Origins and consequences of cognitive fatigue. *Trends in Cognitive Sciences*.

Plack, C. J., Oxenham, A. J., Fay, R. R., & Popper, A. N. (2005). Pitch : neural coding and perception..

Ponsot, E., Arias, P., & Aucouturier, J.-J. (2018). Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America*, *143 1*, EL19.

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, *115*(15), 3972–3977.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *ieee assp magazine*, *3*(1), 4–16.

Rachman, L., Liuni, M., Arias, P., Lind, A. G., Johansson, P., Hall, L., ... Aucouturier, J.-J. (2017). David: An open-source platform for real-time transformation of infra-

segmental emotional cues in running speech. *Behavior Research Methods*, *50*, 323 - 343.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166.

Rosenbek, J. C., Crucian, G. P., Leon, S. A., Hieber, B., Rodriguez, A. D., Holiway, B., ... Gonzalez-Rothi, L. (2004). Novel treatments for expressive aprosodia: A phase i investigation of cognitive linguistic and imitative interventions. *Journal of the International Neuropsychological Society*, *10*, 786 - 793.

Ross, E. D. (1981). The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. *Archives of neurology*, *38*(9), 561–569.

Ross, E. D., Thompson, R., & Yenkosky, J. P. (1997). Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. *Brain and Language*, *56*, 27-54.

Rousseaux, M., Daveluy, W., & Kozlowski, O. (2010). Communication in conversation in stroke patients. *Journal of Neurology*, *257*, 1099-1107.

Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, *109*(4), 597–610.

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P. E. G., & Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biology*, *25*, 3079-3085.

Sandson, J., & Albert, M. L. (1984). Varieties of perseveration. *Neuropsychologia*, *22*(6), 715–732.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*, 2408 - 2412.

Saway, B. F., Palmer, C., Hughes, C., Triano, M., Suresh, R. E., Gilmore, J., ... Rowland, N. C. (2024). The evolution of neuromodulation for chronic stroke: from neuroplasticity mechanisms to brain-computer interfaces. *Neurotherapeutics*, *21*(3), e00337.

Schirmer, A., & Kotz, S. A. (2006, January). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30. doi: 10.1016/j.tics.2005.11.009

Seddoh, S. A. (2002). How discrete or independent are"affective prosody"and"linguistic prosody"? *Aphasiology*, *16*(7), 683–692.

Sergent, C., Corazzol, M., Labouret, G., Stockart, F., Wexler, M., King, J.-R., ... Pressnitzer, D. (2021). Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nature communications*, *12*(1), 1149.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, 639–650.

Shadlen, M. N., & Kiani, R. (2011). Consciousness as a decision to engage. In *Characterizing consciousness: from cognition to the clinic?* (pp. 27–46). Springer.

Sihvonen, A. J., Sammler, D., Ripollés, P., Leo, V., Rodríguez-Fornells, A., Soinila, S., & Särkämö, T. (2021). Right ventral stream damage underlies both poststroke aprosodia and amusia. *European Journal of Neurology*, *29*, 873 - 882.

Spaccavento, S., Marinelli, C. V., Nardulli, R., Macchitella, L., Bivona, U., Piccardi, L., ... Angelelli, P. (2019). Attention deficits in stroke patients: the role of lesion characteristics, time from stroke, and concomitant neuropsychological deficits. *Behavioural neurology*, *2019*(1), 7835710.

Sperber, C., Gallucci, L., Mirman, D., Arnold, M., & Umarova, R. M. (2023). Stroke lesion size–still a useful biomarker for stroke severity and outcome in times of high-dimensional models. *NeuroImage: Clinical*, *40*, 103511.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Stockbridge, M. D., Sheppard, S. M., Keator, L. M., Murray, L. L., Blake, M. L., working group, R. H. D., ... others (2022). Aprosodia subsequent to right hemisphere brain damage: A systematic review and meta-analysis. *Journal of the International Neuropsychological Society*, *28*(7), 709–735.

Sundaram, R. H. (2000). *The baum-welch algorithm* (Tech. Rep.). Technical Report.

Taghia, J., Cai, W., Ryali, S., Kochalka, J., Nicholas, J., Chen, T., & Menon, V. (2018). Uncovering hidden brain state dynamics that regulate performance and decision-

making during cognition. *Nature communications*, *9*(1), 2505.

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, *357*(6353), 797–801.

Tompkins, C. (2012). Rehabilitation for cognitive-communication disorders in right hemisphere brain damage. *Archives of physical medicine and rehabilitation*, *93 1 Suppl*, S61-9.

Tremblay, P., & Dick, A. S. (2016). Broca and wernicke are dead, or moving past the classic model of language neurobiology. *Brain and language*, *162*, 60–71.

Tuppin, P., Rivière, S., Rigault, A., Tala, S., Drouin, J., Pestel, L., ... Fagot-Campagna, A. (2016). Prevalence and economic burden of cardiovascular diseases in france in 2013 according to the national health insurance scheme database. *Archives of cardiovascular diseases*, *109 6-7*, 399-411.

Uekermann, J., Abdel-Hamid, M., Lehmkämper, C., Vollmoeller, W., & Daum, I. (2008). Perception of affective prosody in major depression: A link to executive functions? *Journal of the International Neuropsychological Society*, *14*, 552 - 561.

Ukaegbe, O. C., Holt, B. E., Keator, L. M., Brownell, H., Blake, M. L., Lundgren, K., & Evidence, R. H. D. W. G. (2022). Aprosodia following focal brain damage: What's right and what's left? *American journal of speech-language pathology*, 1-16.

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015). How musical expertise shapes speech perception: evidence from auditory classification images. *Scientific reports*, *5*(1), 14489.

Villain, M., Cosin, C., Glize, B., Berthoz, S., Swendsen, J., Sibon, I., & Mayo, W. (2016). Affective prosody and depression after stroke: A pilot study. *Stroke*, *47*(9), 2397–2400.

Volterra, V. (1930). Theory of functionals and of integral and integro-differential equations.

Vuilleumier, P. (2004). Anosognosia: the neurology of beliefs and uncertainties. *Cortex*, *40*(1), 9–17.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *Isca tutorial and research workshop (itrw) on speech and emotion* (pp. 5–7).

Wiener, N. (1966). *Nonlinear problems in random theory.*

Worrall, L., Brown, K., Cruice, M., Davidson, B., Hersh, D., Howe, T., & Sherratt, S. (2010). The evidence for a life-coaching approach to aphasia. *Aphasiology*, *24*(4), 497–514.

Wylie, G. R., Yao, B., Sandry, J., & DeLuca, J. (2021). Using signal detection theory to better understand cognitive fatigue. *Frontiers in psychology*, *11*, 579188.

Xie, R., & Xu, L. (2024). The neural mechanisms behind mind wandering phenomenon: A brief review. *assessment*, *6*(4), 82–85.

Xuan, G., Zhang, W., & Chai, P. (2001). Em algorithms of gaussian mixture model and hidden markov model. *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, *1*, 145-148 vol.1.

Yu, S., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden markov model. *IEEE Signal Processing Letters*, *10*, 11-14.

Yuan, M.-z., Li, F., Fang, Q., Wang, W., Peng, J.-j., Qin, D.-y., ... Liu, G.-w. (2018). Research on the cause of death for severe stroke patients. *Journal of clinical nursing*, *27*(1-2), 450–460.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*, 37-46.

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256 5058*, 846-9.

Zhang, C., & Kool, W. (2024). Inferring mind wandering from perceptual decision making. *OSF preprint https://osf.io/mxtbh.*

List of figures

List of tables

List of Abbreviations

2A/IFC	Two-Alternative/Interval Forced Choice
AIRTAC2	Aide informatisée pour la rééducation des troubles auditifs centraux 2
BDAE	Boston Diagnostic Aphasia Examination
CI	Classification Images, or Confidence Interval
ENG	Engaged state
GLM	Generalized Linear Model
HADS	Hospital Anxiety and Depression scale
HMM	Hidden Markov Model
IFG	Inferior Frontal Gyrus
JND	Just Noticeable Difference
LAMA	Logiciel d'Attention en Modalité Auditive
MAP	Maximum a posteriori
MBEA	Montreal Battery of Evaluation of Amusia
MEC	Montreal Evaluation of Communication
MLE	Maximum-likelihood estimation
PER	Perseverative state
RE	Relative error
RH	Right hemisphere
RMSE	Root Mean Square Error
RT	Response Time
SDT	Signal detection theory

Résumé Cette thèse vise à améliorer le diagnostic et la compréhension des déficits de perception de la prosodie chez les patients ayant subi un AVC dans l'hémisphère droit, en utilisant la technique psychophysique de la corrélation inverse. Alors que les outils d'évaluation standard manquent de sensibilité et de précision pour détecter de telles déficiences, la corrélation inverse offre un moyen de quantifier les représentations sensorielles internes et le bruit interne des patients, fournissant un aperçu des mécanismes cognitifs et perceptifs perturbés par l'accident vasculaire cérébral.

Dans la première partie, nous avons appliqué des paradigmes de corrélation inverse et des méthodes d'analyse classiques à des données comportementales provenant de patients ayant subi un accident vasculaire cérébral et de témoins sains. Ce travail initial a révélé des profils pathologiques distincts et des différences dans les noyaux perceptifs et le bruit interne. Cependant, deux limitations critiques sont apparues : le faible nombre d'essais disponibles et la présence de persévération, une réponse répétitive qui déforme les correspondances entre la réponse et le stimulus. Ces problèmes ont compromis la fiabilité des techniques standard de corrélation inverse et ont motivé le développement d'approches plus robustes.

Dans la partie II, nous avons présenté plusieurs contributions méthodologiques. Trois nouvelles techniques ont été développées pour estimer le bruit interne sans nécessiter d'essais à double passage, y compris une méthode basée sur des intervalles de confiance dérivés de GLM. Ces méthodes sont plus précises que les approches classiques, en particulier dans les scénarios à faible nombre d'essais. Nous avons également proposé un GLM-HMM pour estimer conjointement les noyaux perceptifs et identifier les états internes latents, tels que l'engagement et la persévération, sur la base de l'intégration du stimulus et des schémas de réponse.

Dans la troisième partie, nous avons appliqué ces outils perfectionnés à des données cliniques. Le GLM-HMM a révélé qu'une grande partie du bruit interne précédemment mesuré à l'aide des techniques classiques était gonflée par le comportement persévératif. Il

est important de noter que les estimations du bruit dérivées des nouvelles méthodes sont en corrélation avec des évaluations cliniques distinctes, ce qui indique une amélioration de la validité de la construction. Les probabilités de transition entre les états latents sont apparues comme de nouveaux biomarqueurs potentiels : les patients ayant subi un accident vasculaire cérébral n'étaient pas plus susceptibles d'entrer dans des états persévératifs que les témoins, mais ils présentaient une capacité réduite à s'en remettre. En outre, la persévération chez les patients était associée à des temps de réponse plus lents et, avant la persévération, à des séquences plus longues de réponses identiques, ce qui suggère un mécanisme cognitif différent de celui des témoins. Notamment, le retour à l'engagement tend à coïncider avec des essais de moindre difficulté présentantun meilleur alignement entre le stimulus et le noyau interne du patient.

Mots-clés : AVC de l'hémisphère droit, perception de la prosodie, corrélation inverse, bruit interne, persévération, GLM-HMM

Abstract *(in english)* **Title: Reverse-correlation modeling of deficits of prosody perception in right-hemisphere stroke**

This thesis aims to improve the diagnosis and understanding of prosody perception deficits in right-hemisphere stroke patients using the psychophysical technique of reverse correlation. While standard assessment tools lack the sensitivity and precision to detect such impairments, reverse correlation offers a means to quantify patients' internal sensory representations and internal noise, providing insight into the cognitive and perceptual mechanisms disrupted by stroke.

In Part I, we applied reverse correlation paradigms and classical analysis methods to behavioral data from stroke patients and healthy controls. This initial work revealed distinct pathological profiles and differences in perceptual kernels and internal noise. However, two critical limitations emerged: the low number of available trials and the presence of perseveration, repetitive responding that distorts response-stimulus mappings. These issues compromised the reliability of standard reverse-correlation techniques and motivated the development of more robust approaches.

In Part II, we introduced several methodological contributions. Three novel techniques were developed to estimate internal noise without requiring double-pass trials, including a method based on GLM-derived confidence intervals. These methods were more accurate than classical approaches, particularly in low-trial scenarios. We also proposed a GLM-HMM to jointly estimate perceptual kernels and identify latent internal states, such as engagement and perseveration, based on both stimulus integration and response patterns.

In Part III, we applied these refined tools to clinical data. The GLM-HMM revealed that much of the internal noise previously measured using classical techniques was inflated by perseverative behavior. Importantly, noise estimates derived from the new methods correlated with distinct clinical assessments, indicating improved construct validity. Transition probabilities between latent states emerged as new potential biomarkers: stroke patients were not more likely to enter perseverative states than controls but exhibited

reduced ability to recover from them. Moreover, perseveration in patients was associated with slower response times and, before them, longer sequences of identical responses, suggesting a different cognitive mechanism than in controls. Notably, switches back to engagement tended to coincide with trials of lower difficulty / greater alignment between the stimulus and the patient's internal kernel.

Keywords: Right-hemisphere stroke, Prosody perception, Reverse correlation, Internal noise, Perseveration, GLM-HMM